Computational analysis of the immunogenicity and sequence diversity of *Mycobacterium tuberculosis* PPE_MPTR proteins

by
Antoinette Danielle Colic

Thesis presented in partial fulfilment of the requirements
for the degree of Master of Science (Molecular Biology) in the
Faculty of Medicine and Health Sciences
at Stellenbosch University

Supervisor: Prof Samantha Sampson
Co-supervisor: Prof Alan Christoffels

March 2017

Declaration

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Date: March 2017

Abstract

Mycobacterium tuberculosis presents a substantial health risk to humans, particularly in Africa. Prevention of infectious diseases via vaccination is the most effective strategy in decreasing prevalence; however the current BCG vaccine against tuberculosis has shown varying levels of efficacy. *M. tuberculosis* infection represents an on-going interaction between the host and the bacteria, of which we do not yet fully understand all the mechanisms contributing to the pathogenesis at a molecular level. A deeper understanding of host-pathogen interactions is an important step towards developing new and more effective vaccines and therefore combating the disease. Protective immunity against *M. tuberculosis* is induced by stimulating antigen specific T-cells, which recognise peptide antigens presented by HLA molecules on infected cells. Identifying epitopes that are capable of binding to HLA molecules and eliciting T-cell responses form part of the development of subunit vaccines.

An area of mycobacterial biology that is poorly understood is the function of the PE/PPE proteins. These proteins are a large, genetically diverse family of immunogenic proteins that are predicted to play a role in modulating host immune responses. In particular, the PPE major polymorphic tandem repeat (PPE_MPTR) proteins are a subgroup of the PE/PPE proteins which are restricted to pathogenic mycobacterial species and represent one of the most genetically diverse set of proteins within the *M. tuberculosis* proteome. While many studies have investigated the presence of T-cell epitopes within the PE/PPE family of proteins, no studies have focused specifically on the PPE_MPTR subfamily. Based on the extreme variation in both the length and genetic diversity of the PPE_MPTR proteins, it has been speculated that they may represent a source of antigenic variation which allows the organism to escape antigen-specific host responses. Given the hyper-variable nature of the PPE_MPTR proteins and their possible role in host-pathogen interactions, genetic diversity within the PPE_MPTR proteins may differentially modulate human immune response. Furthermore, epitopes within the PPE_MPTR proteins may be possible subunit vaccine candidates for *M. tuberculosis.*

Conventional experimental techniques used to identify potential epitopes can often be time consuming and expensive. Various computational tools exist to predict binding of peptide sequences to various HLA alleles. Using a collection of known *M. tuberculosis* epitopes from the Immune Epitope Database (IEDB), an evaluation of the current open source HLA class II prediction tools has been performed, with the results used to inform an *in silico* identification of human CD4+ T-cell epitopes within the PPE_MPTR proteins. Characterisation of the genetic diversity of these proteins is also an essential step in improving our understanding of this protein family. Publically available whole genome sequence data from strains belonging to various lineages has been used to investigate the level of sequence diversity within these *ppe_mptr* genes, and the impact of genetic variants on epitope density has been investigated. To date, this study is the most comprehensive analysis of the genetic variation of the *ppe_mptr* genes. Predicted epitopes have been filtered using a reverse vaccinology approach in order to identify possible subunit vaccine candidates for *M. tuberculosis*.

Findings from epitope prediction analysis support the hypothesis of host-pathogen interactions for the PPE_MPTR proteins. Genetic variation results indicate that certain PPE_MPTR proteins are highly variable while others are relatively conserved across strains, and that genetically diverse regions are less likely to contain epitopes. Therefore no evidence to support antigenic variation was found. Areas of high and low epitope density are correlated to areas of non-repeat and repeat regions within the genome respectively, and therefore epitopes within the PPE_MPTR proteins are conserved non-repeating peptides. This is consistent with previous literature on the conservation of reported *M. tuberculosis* epitopes within clinical strains. Further studies are therefore needed to determine the role of the variable copy number of repeats found within the PPE_MPTR proteins. Possible vaccine candidates with high predicted population coverage in African countries within the PPE_MPTR proteins have been identified.

Opsomming

Mycobacterium tuberculosis bied ‘n aansienlike gesondheidsrisiko vir mense, veral in Afrika. Vroegtydige inenting is die mees suksesvolle strategie in die bekamping van aansteeklike siektes. Ongelukkig het BCG *(Bacillus Calmette–Guérin)*, die enigste entstof teen tuberkulose, wisselvallige sukses bereik in die taak. *M. tuberculosis* infeksies verteenwoordig ‘n aanhoudende stryd tussen beide gasheer en bakterie, waarvan die molekulêre meganismes wat bydrae tot patogenese nog nie volledig beskryf is nie. Dus, 'n meer indiepte begrip van die gasheer-patogeen interaksie sal die ontwikkeling van doeltreffende inentingsstowwe bevorder. Beskermende immuniteit teen *M. tuberculosis* word geïnduseer deur die stimulering van antigeen spesifieke T-selle wat peptied antigene herken wat deur HLA (human leucocyte antigen) molekules blootgestel word. Die identifisering van epitope wat aan HLA molekules bind en T-sel reaksies lok, vorm deel van die ontwikkeling van subeenheid entstowwe.

Die bydrae en funksie van PE/PPE proteïene in mikobakteriële biologie word tans nog nie volledig verstaan nie. PE/PPE proteïene is afkomstig van ‘n groot, geneties diverse familie van immunogeniese proteïene waarvan die rol in modulering van die gasheer immuun respons voorspel word. Die PPE “major polymorphic tandem repeat” (PPE_MPTR) proteïene, wat ‘n subgroep van die PE/PPE proteïene vorm, is beperk to die patogeniese mikobakteriële spesies en verteenwoordig die mees genetiese diverse stel van proteïene in die *M. tuberculosis* proteoom. Alhoewel baie navorsing al uitgevoer is oor die teenwoordigheid van T-sel epitope binne die konteks van die PE/PPE familie van proteïene, is daar nog geen studie wat spesifiek fokus op die PPE_MPTR subfamilie nie. As gevolg van die hoë variasie in beide die lengte en genetiese diversiteit van PPE_MPTR proteïene, word daar gespekuleer dat PPE_MPTR ‘n bron van antigeniese variasie is wat die organisme in staat stel om die antigeen-spesifieke gasheer respons te vermy. Die hoë variasie van PPE_MPTR proteïene en hul moontlike rol in gasheer-patogeen interaksie kan die gasheer immunrespons moduleer. Epitope binne die PPE_MPTR proteïene kan dus goeie kandidate vir subeenheid entstowwe teen *M. tuberculosis* wees.

Tradisionele metodes wat gebruik word om potensiële epitope te identifiseer is dikwels tydrowend en duur. Dus, rekenaargebaseerde tegnieke was ontwikkel om die binding van peptiede aan verskeie HLA allele te voorspel. Verskeie oopbron HLA klass II voorspellings tegnieke was gebruik om $CD4^+$ T-sel epitope *in silico* te identifiseer binne die PPE_MPTR proteïene, deur gebruik te maak van ‘n versameling bekende *M. tuberculosis* epitope wat verkry is vanaf die Immune Epitope Databasis (IEDB). Die karakterisering van PPE_MPTR proteïene is ‘n noodsaaklike stap in die ontwikkeling van kennis oor hierdie proteïen familie. Openbare heel genoom data van stamme, wat aan verskeie stamfamilies behoort, was gebruik om variasie te bepaal binne die *ppe-mtpr* gene, asook die impak van genetiese variante op epitoop digtheid was bepaal. Die huidige studie, wat die PPE_MPTR genetiese variasie ondersoek, is die mees omvattendste analise tot dusver. Voorspelde epitope was geselekteer deur gebruik te maak van ‘n tru-vaksinologie benadering om subeenheid entstowwe teen *M. tuberculosis* te identifiseer.

PPE_MPTR protein epitoop voorspellings staaf die hipotese van gasheer-patogeen interaksie. Analises rakende die genetiese variasie dui daarop dat sekere PPE_MPTR proteïene baie veranderlik voorkom, terwyl ander relatief behoue bly binne hul verskillende stamme. Dus, geen bewyse wat antigeniese varisasie staaf was gevind nie. Areas met hoë en lae epitoop digtheid onderskeidelik, korreleer goed met nie-herhalende en herhalende dele binne die genoom. Dus, epitope binne die PPE_MPTR proteïene is konservatiewe, nie-herhalende peptiede. Hierdie is in lyn met vorige literatuur wat die bewaring van *M. tuberculosis* epitope binne kliniese stamme aandui. Verdere navorsing is nodig om die rol van die variasie in aantal herhalings binne die PPE_MPTR proteïene te bepaal. Entstof kandidate wat hoë voorspelde dekking bied onder die Afrika lande is geïdentifiseer binne PPE_MPTR proteïene.

Acknowledgements

I would like to say a sincere thank you to the following people, without all of your support this thesis would not have been possible:

- To my supervisor, Prof Samantha Sampson. I feel extremely privileged to have had the pleasure of meeting and working with you. Your love for science is truly inspiring. Thank you for always making the time to brainstorm about ideas and critically evaluate my work. Thank you for all of the support this year, when at times I did not think it would be possible to keep going, your words of encouragement were what pushed me to persevere.

- To my co-supervisor Alan Christoffels, as well as to Prof Gerhard Tromp, Prof Luisa Azevedo, Dr Anzaan Dippenaar, and Dr Ruben van der Merwe for your willingness to assist on the technical aspects of this work.

- To Prof Paul van Helden and the Division of Molecular Biology and Human Genetics, and in particular to the Mycobactomics lab for hosting me and making me feel welcome.

- To my parents, Johan and Brenda Niehaus, to my sister Jacqui, to Mladen and all of my other friends and family. I can't thank you enough for the support and for pushing me to be the best version of myself possible. Thank you for starting this journey with me and sticking with me until the end.

I would like to thank the National Research Foundations (NRF) for the financial support.

Table of Contents

Declaration i
Abstract ii
Opsomming iii
Acknowledgements iv
Table of Contents v
List of Abbreviations ix
Chapter 1: General Introduction 1
1.1 Background 2
1.1.1 Tuberculosis and BCG vaccine 2
1.1.2 Host-pathogen interaction 2
1.1.3 Immunity against TB 2
1.1.4 Current approaches to new TB vaccine development 3
1.1.5 PE/PPE proteins 4
1.1.6 PPE_MPTR proteins and their possible interaction with the immune system 4
1.2 Problem Statement 5
1.3 Hypotheses 6
1.4 Aims and Objectives 6
1.4.1 Aim 1: *In silico* identification of human CD4+ T-cell epitopes in PPE_MPTR proteins 6
1.4.1.1 Objective 1: Determination of the optimal epitope prediction pipeline 6
1.4.1.2 Objective 2: Prediction of PPE_MPTR epitopes 7
1.4.2 Aim 2: Characterisation of the level of genetic diversity within the PPE_MPTR family members 7
1.4.3 Aim 3: Determine the impact of genetic variants on epitope prediction 7
1.4.4 Aim 4: Identification of potential vaccine candidates 7
1.5 Significance of Research 7
1.6 Scope and Limitations 7
1.7 Chapter Overview 8
1.8 References 9
Chapter 2: Literature Review: 12
2.1 Introduction 13
2.1.1 The search for a new vaccine against Tuberculosis 13
2.1.2 Immunoinformatics 14

2.1.3 Adaption of RV workflows to *M. tuberculosis* 16
2.2 Immunological Databases 17
2.2.1 T-Cell epitope databases 17
2.2.2 Human immune cell databases 18
2.2.3 Mycobacterium specific databases 18
2.2.4 Data repositories for machine learning 19
2.3 T-Cell Epitope Prediction 19
2.3.1 Prediction algorithms 19
2.3.2 Data quantity and quality 20
2.3.3 Epitope prediction tools 21
2.3.4 HLA class I antigen presentation pathway tools 23
2.4 Protective Coverage of Vaccine Candidates 23
2.4.1 HLA alleles 23
2.4.2 Prediction of potential population coverage 24
2.5 Identification of Potential TB Vaccine Candidates 25
2.5.1 Selection criteria: family of proteins 25
2.5.2 Selection criteria: biological function 26
2.5.3 Whole genome approaches 27
2.5.4 Protective coverage: 27
2.5.5 Experimental validation: 27
2.6 Conclusion 28
2.7 References 29
Chapter 3: Evaluation of MHC II Epitope Prediction Tools 36
3.1 Introduction 37
3.2 Methods 37
3.2.1 Epitope data 37
3.2.2 Tools evaluated 39
3.2.3 Interpretation of prediction results 40
3.2.4 Accuracy by HLA allele 41
3.3 Results 41
3.3.1 Overall accuracy, sensitivity and specificity 41
3.3.2 Results per HLA allele 42
3.4 Conclusion 45
3.5 References 46
Chapter 4: MHC Class II Epitope Prediction 47
4.1 Introduction 48

4.2 Methods 48
4.2.1 PPE_MPTR protein sequence data 48
4.2.2 HLA alleles 49
4.2.3 Prediction of epitopes 49
4.2.4 Determination of the PPE boundary 51
4.2.5 Identification of promiscuous epitopes 51
4.2.6 Binding ability of HLA alleles 51
4.3 Results 52
4.3.1 Number of epitopes in PPE versus MPTR region 52
4.3.2 Patterns of fluctuation along the length of the protein 53
4.3.3 Promiscuous epitopes 54
4.3.4 Binding ability of HLA alleles 56
4.4 Conclusion 57
4.5 References 58
4.6 Appendices 59
A: Proportion of epitopes versus non-epitopes, unique binders versus promiscuous epitopes, and distribution of promiscuous epitopes. 59
B: Patterns of predicted epitopes showing distribution of promiscuity along length of each protein. 66
C: Binding ability of HLA alleles per PPE_MPTR protein 69
Chapter 5: Genetic Diversity of the PPE_MPTR Proteins 80
5.1 Introduction 81
5.1.1 PPE_MPTR genomics 81
5.1.2 Determining the genetic diversity of the PPE_MPTR proteins 81
5.1.3 Current approach used 82
5.2 Methodology 83
5.2.1 Methodology summary 83
5.2.2 Strain data 84
5.2.3 Ortholog coordinates 87
5.2.4 IS*6110* insertion elements 88
5.2.5 Tandem repeats 88
5.2.6 Determination of variants 88
5.2.7 Effect of micro-mutations on epitope prediction 89
5.2.8 Epitope density within repeat versus non-repeat regions 90
5.3 Results 91
5.3.1 Summary of genetic variation 91
5.3.2 IS*6110* insertion elements 92

5.3.3 Effect of micro-mutations on epitope prediction 92
5.3.4 Epitope density within repeat versus non-repeat regions 97
5.4 Conclusion 101
5.5 References 103
Chapter 6: Identification of Potential Vaccine Candidates 105
6.1 Introduction 106
6.2 Methods 107
6.3 Results 108
6.3.1 Results summary 108
6.3.2 Epitope cluster analysis 108
6.3.3 Potential population coverage 109
6.4 Conclusion 111
6.5 References 112
6.6. Appendix 113
A: Population coverage for high burden TB countries in Africa 113

List of Abbreviations

α:	alpha
γ:	gamma
π:	nucleotide diversity
aa:	Amino Acid
AFND:	Allele Frequency Net Database
ANN:	Artificial Neural Network
ARB:	Average Relative Binding
BCG:	Bacillus Calmette-Guérin
bp:	Base Pair
CASTB:	Comprehensive analysis server for TB
DFRMLI:	Dana-Farber Repository for Machine Learning in Immunology
HIV:	Human Immunodeficiency Virus
HLA:	Human Leukocyte Antigen
HMM:	Hidden Markov Models
IC50:	Half maximal inhibitory concentration
IEDB:	Immune Epitope Database
IFN:	Interferon
IL:	Interleukin
IMGT:	International ImMunoGeneTics
Indel:	Insertions and Deletions
M. tuberculosis:	*Mycobacterium tuberculosis*
MEGA:	Molecular Evolutionary Genetics Analysis
MenB:	Serogroup B Meningococcus
MHC:	Major Histocompatibility Complex
MPTR:	Major Polymorphic Tandem Repeats
NCBI:	National Center for Biotechnology Information
nM:	Nanomolar
NN:	Neural Network
nsSNP:	Non-Synonymous SNP
PE:	Proline-Glutamate
PGRS:	Polymorphic GC-rich Sequence
PLS:	Partial Least Squares
PPE:	Proline-Proline-Glutamate
PSSM:	Position Specific Scoring Matrices
QSAR:	Quantitative Structure-Activity Relationship
RV:	Reverse Vaccinology
SDR:	Specificity-Determining Residues
SMM:	Stabilization Matrix Alignment
SNP:	Single Nucleotide Polymorphisms
sSNP:	Synonymous SNP
SVM:	Support Vector Machines
TAP:	Transfer-Associated Protein
TB:	Tuberculosis
TBDB:	TB Database
TLR:	Toll-Like Receptor
TNF:	Tumor necrosis factor
TRF:	Tandem Repeat Finder
VNTR:	Variable Number of Tandem Repeats
WGS:	Whole Genome Sequence
WHO:	World Health Organisation

Chapter 1: General Introduction

1.1 Background

1.1.1 Tuberculosis and BCG vaccine

Mycobacterium tuberculosis, the causative agent of tuberculosis (TB) presents a substantial health risk to humans, particularly in Africa. An estimated 9.6 million people developed TB and 1.5 million died from the disease in 2014 worldwide according to the World Health Organisation (WHO) (WHO 2015). Despite the fact that a vaccine and a drug regime for TB exist, with certain countries obtaining near eradication, TB still remains a threat worldwide. South Africa in particular is ranked as one of the 30 high burden countries in the world, along with 15 other African countries, including Angola, Congo, Central African Republic, DR Congo, Ethiopia, Kenya, Lesotho, Liberia, Mozambique, Namibia, Nigeria, Sierra Leone, the United Republic of Tanzania, Zambia and Zimbabwe.

Prevention of infectious diseases via vaccination is still the most effective strategy in decreasing prevalence and minimizing the impact of disease on the human population (Delany *et al*. 2014). The Bacillus Calmette-Guérin (BCG) vaccine, developed more than 80 years ago is still the only approved vaccine against TB; however it does not prevent pulmonary disease in adults in all cases and has shown varying levels of efficacy (WHO 2015). In a review and meta-analysis of the efficacy and duration of protection by BCG, it was found that protection against pulmonary TB in adults is highly variable with substantial protection reported in certain studies but with absence of clinically important benefits in others, and duration of protection reported at an average of up to only 10 years (Abubakar *et al*. 2013). Despite numerous efforts to develop additional, more effective vaccines against *M. tuberculosis*, we have yet to find one.

1.1.2 Host-pathogen interaction

M. tuberculosis infection represents an on-going interaction between the host and the bacteria, of which we do not yet fully comprehend all the mechanisms contributing to the pathogenesis at a molecular level. We are yet to gain a full understanding of how *M. tuberculosis* manages to successfully evade the human immune system, and create a niche environment. Investigating the virulence and pathogenesis of *M. tuberculosis* is an important step towards developing new and more effective vaccines and therefore combating the disease. The genome of *M. tuberculosis* contains 4090 genes, of which 3933 encode proteins; however approximately a quarter of coding sequences do not have an assigned functional category according to TubercuList (a relational database presenting genome-derived information about the reference strain *M. tuberculosis* H37Rv) (Lew *et al*. 2011). Gaining a more comprehensive knowledge of the function *M. tuberculosis* proteins play during host-pathogen interaction is an important first step towards identifying potential vaccine targets.

1.1.3 Immunity against TB

Innate immune response against *M. tuberculosis* following inhalation into the lungs primarily includes ingestion by alveolar macrophages, which initiates a number of signalling pathways involved in inflammation as well as the development and regulation of the adaptive immune response (Philips & Ernst 2012). Past literature on the adaptive immune response to *M. tuberculosis* has focused on the role of T-cells rather than the role of B-cells and antibodies (which remain relatively unclear) (Kozakiewicz *et al*. 2013). The crucial role of T-cells in the adaptive immune response to TB has been demonstrated in seminal studies using mouse models, where transgenic mice with deficient major histocompatibility complex (MHC) class II molecules or CD4+ T-cells showed increased susceptibility to *M. tuberculosis* when compared with wild-type mice (Caruso *et al*. 1999). The role of CD4+ T-cells is also evident when investigating human immunodeficiency virus (HIV) positive patients who show increased susceptibility to TB when CD4+ cell counts are low (Kwan & Ernst 2011). Mouse models have also highlighted the importance of CD8+ T-cells, where mutant mice which lack MHC class I molecules and fail to develop functional CD8+ T-cells show a higher bacterial load and a decreased time to death compared to wild-type mice (Flynn *et al*. 1992).

Protective immunity against *M. tuberculosis* is induced in humans by stimulating antigen specific T-cells, which recognise peptide antigens presented by human leukocyte antigen (HLA) molecules on infected cells. CD8+ T-cells recognise antigens presented by HLA class I molecules, while CD4+ T-cells recognise antigens presented by HLA class II molecules. Peptides presented by HLA class I molecules are mainly derived from the degradation of intracellular proteins whereas peptides presented by HLA class II molecules are derived from the degradation of external proteins which have been internalized by antigen presenting cells (such as macrophages and dendritic cells) (Murphy 2011). Therefore while CD8+ T-cells play an important role in the adaptive immune response against *M. tuberculosis*, the activation of CD4+ T-helper cells is critical in regulating the adaptive immune response. Any potential vaccine candidate must therefore be able to stimulate CD4+ T-lymphocytes.

1.1.4 Current approaches to new TB vaccine development

The current approach to TB vaccination uses BCG, a live attenuated vaccine derived from bovine mycobacteria, which has lost the ability to cause disease in humans but is still similar enough to *M. tuberculosis* to induce protective immunity (Luca & Mihaescu 2013). In addition to live attenuated vaccines, new vaccine candidates currently under development include subunit vaccines and mycobacterial lysates. Post-exposure vaccines which aim to prevent reactivation and relapse, and therapeutic vaccines which can be used to supplement drug regimens, are also under extensive research (Loxton *et al*. 2016).

Subunit or peptide based vaccines consist of selected mycobacterial antigens which are delivered via viral vectors or adjuvants (Loxton *et al*. 2016). Immune memory does not need to be generated for every peptide in a pathogen, and T-cell responses to one or a set of peptides can be sufficient to provide protective immunity (Keating & Noble 2003; Song *et al*. 2015). Subunit vaccines offer the advantage of decreased possible adverse reactions since a vaccine would only contain essential antigens, as well as the ability to produce sufficient quantities of purified antigenic components as opposed to other vaccination approaches (Nascimento & Leite 2012). In addition, recent literature has shown the possibility of enhancing the immunity induced by live attenuated candidates through the addition or overexpression of promising antigens (McShane *et al*. 2005; Billeskov *et al*. 2012). Examples of subunit vaccine candidates currently undergoing clinical trials are shown in Table 1.

Table 1.1: Subunit vaccine candidates in current clinical trials. There are currently 7 subunit vaccines undergoing clinical trials. The specific *M. tuberculosis* antigens within each is shown (Loxton *et al*. 2016).

Name	Delivery Method	Clinical Phase	*M. tuberculosis* antigens included
ID93	Adjuvant (GLA-SE)	I	Rv2608 (PPE42), Rv3619 (EsxV), Rv3620 (EsxW), Rv1813 (latency-associated protein)
H1	Adjuvant (CAF01/IC31)	I	Ag85B (mycolytransferase) and ESAT-6
H4	Adjuvant (IC31)	I	TB10.4 (Rv0288, member of ESAT6 family) and Ag85B
H56	Adjuvant (IC31)	I	Ag85B, ESAT-6 and Rv2660c
M72A	Adjuvant (AS01)	II	MTB32A (PPE18), and MTB39A (serine protease)
MVA85A	Viral (Modified Vaccinia Ankara vector)	II	Ag85A (mycolytransferase)
Aeras402/ Crucell Ad35	Viral (adenovirus vector)	II	TB10.4, Ag85A and Ag85B.

Bacterial peptides that are recognised by the host immune system are called epitopes, and identifying epitopes that are capable of binding to HLA molecules and eliciting T-cell responses form part of the development of subunit vaccines. Experimental techniques used to identify potential epitopes can often be time consuming and expensive. Reverse vaccinology (RV) pipelines which use *in silico* approaches to identify potential subunit vaccine candidates provide an alternative option. RV pipelines make use of genomic data, sequence analysis tools and predictive algorithms to search for potential vaccine candidates computationally. These approaches may offer the benefit of identifying antigens which may be difficult to

work with experimentally and therefore missed by traditional approaches. The use of RV approaches in the search for new TB vaccines is expanded upon in Chapter 2.

Two of the subunit vaccine candidates in Table 1.1, ID93 and M72A include antigens from the PE/PPE protein family of *M. tuberculosis*. ID93 which includes PPE42/Rv2608 has shown to be well tolerated and protective in animal models (Bertholet *et al*. 2010) and is currently undergoing phase I clinical trials. M72A, which includes PPE18/Rv1196, has undergone both phase I and II clinical trials and has demonstrated safety as well as both humoral and cell-mediated immune responses (Day *et al*. 2013).

1.1.5 PE/PPE proteins

An area of mycobacterial biology that is poorly understood is the function of the PE/PPE proteins. These proteins are a large, genetically diverse family of immunogenic proteins that are predicted to play a role in modulating host immune responses. While the functions of these proteins are largely unknown (Lew *et al*. 2011), there is evidence for certain members of the family to support interaction with toll-like receptors (TLR) (Basu *et al*. 2007), promote macrophage uptake (Brennan *et al*. 2001), as well as inhibiting antigen processing (Cole *et al*. 1998). Certain members have also been shown to illicit B-cell responses, CD4+ and CD8+ T-cell responses (Sampson 2011).

The genes encoding the PE/PPE proteins make up approximately 10% of the *M. tuberculosis* genome (Cole *et al*. 1998). The PE/PPE proteins are divided into two groups and are named for their proline-glutamate residues (PE) or proline-proline-glutamate (PPE) residues at the start of the encoded proteins. These two groups are then further subdivided based on their C-terminal domains which are highly variable in their length and sequence features (Figure 1.1).

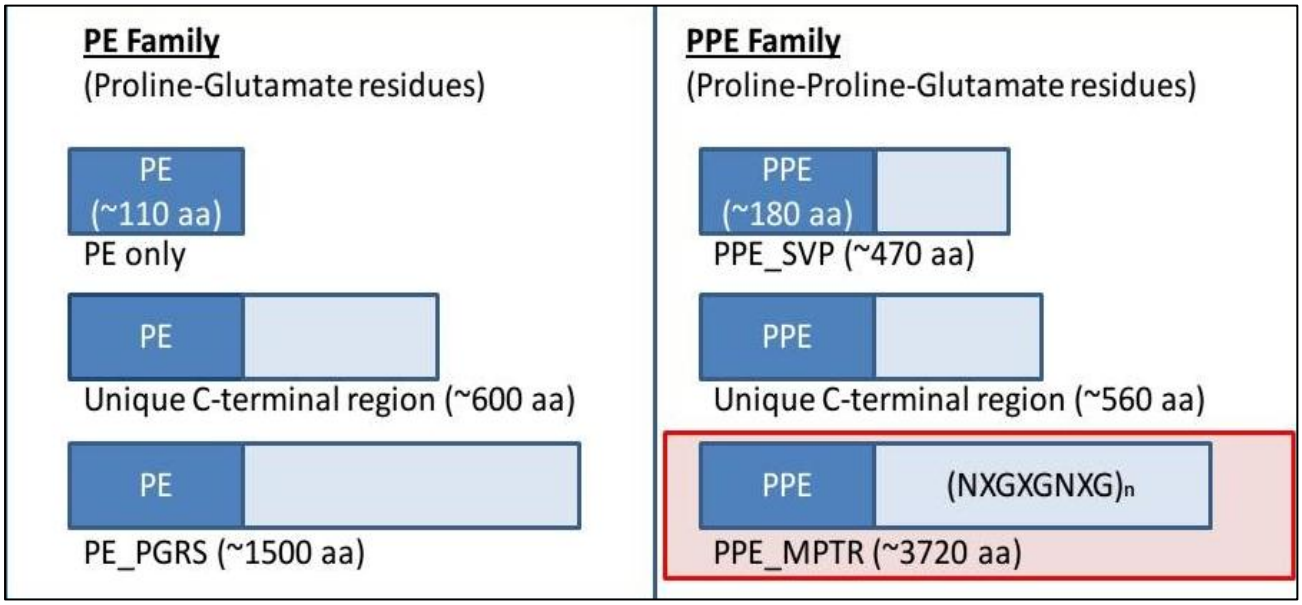


Figure 1.1: PE/PPE Genomics. PE/PPE proteins can be divided into subgroups based on their unique C-terminal features. While the PE/PPE region of the sequence has been shown to be highly conserved between strains, the variable C-terminal domains are highly polymorphic; most notably the polymorphic GC-rich sequence (PGRS) and major polymorphic tandem repeat (MPTR) sub-families which represent two of the most variable regions in the *M. tuberculosis* genome. (Adapted from Sampson (2011)).

1.1.6 PPE_MPTR proteins and their possible interaction with the immune system

The PPE Major polymorphic tandem repeat (MPTR) proteins are a subgroup of the PE/PPE proteins which are restricted to pathogenic mycobacterial species (Gey van Pittius *et al*. 2006). The protein sequence includes the conserved PPE region of roughly 170-180 amino acids, and the MPTR region which ranges from between 420 to 3700 amino acids. The genes encoding the PPE_MPTR proteins are associated with imperfect repeats (Figure 1.1) and are hotspots for recombination events and mutations (Cole *et al*. 1998; McEvoy *et al*. 2009). Evidence indicates that certain PPE_MPTR proteins (PPE_MPTR34 (Rv1917c) and

PPE_MPTR64 (Rv3558)) localise to the cell wall (Sampson *et al*. 2001; Sani *et al*. 2010), placing them in an ideal position to interact with host components and potentially modulate immune responses. PPE_MPTR42 (Rv2608), one of the antigens included in vaccine candidate ID93, has shown to elicit B-cell responses (Choudhary *et al*. 2004; Ireton *et al*. 2010), as well as T-cell responses which confers protective immunity (Bertholet *et al*. 2008). Other studies suggest that PPE_MPTR proteins play multiple roles in infection (Sampson 2011), including blocking trafficking of *M. tuberculosis* to acidified phagosomes (PPE_MPTR10 (Rv0442c), PPE_MPTR16 (Rv1135c) and PPE_MPTR54 (Rv3343c)) (Stewart *et al*. 2005; Brodin *et al*. 2010), interacting with TLR-2 and inducing dendritic cell maturation (PPE_MPTR34 (Rv1917c)) (Bansal *et al*. 2010).

Based on the extreme variation in both the length and genetic diversity of the PPE_MPTR proteins, it has been speculated that they may represent a source of antigenic variation which allows the organism to escape antigen-specific host responses (Karboul et al. 2008; Akhter et al. 2012). Antigenic variation, which is described as the asynchronous expression of functionally conserved yet antigenetically distinct proteins is a mechanism used by many bacterial and non-bacterial pathogens to evade the host immune system (van der Woude & Baumler 2004). Typically, a pathogen will possess the genetic information to produce a family of antigenic variants but only one variant is expressed at a given time, allowing the pathogen to evade or re-infect a host as the variable antigens will no longer be recognised by the host immune system. In general, antigenic variation is achieved through the assortment and recombination of repeated genes or gene segments (Borst 1991). Classic examples of bacterial species in which antigenic variation occur include: *Borrelia burgdorferi* which uses recombination of its *vlsE* genes to produce variable lipoproteins resulting in waves of sickness each time a new variant emerges (Koomey 1997); *Campylobacter fetus* which uses recombination of its SapA genes to produce variable surface layer proteins (Dworkin & Blaser 1997); and *Neisseria meningitidis* and *Neisseria gonorrhoeae* which are able to produce millions of different, antigenically distinct fimbriae/pili that extend from the cell surface (Nassif *et al*. 1993; Hagblom *et al*. 1985). Epitopes within these antigens are generally under selective pressure leading to more variation within the corresponding gene sequences. To date however, *M. tuberculosis* epitopes have been found to be highly conserved (inconsistent with antigenic variation) (Comas *et al*. 2010), and instead the bacteria seems to use the host immune cells to create a niche growing environment (Ernst 2012).

Previous studies have investigated the possibility of antigenic variation within the highly diverse PE_PGRS sub-family (Copin *et al*. 2014). Along with the PPE_MPTR sub-family, these proteins are some of the most variable regions in the *M. tuberculosis* genome (Cole *et al*. 1998). Bioinformatics methods were used to test whether sequence diversity in the PE_PGRS genes may be selected by human T-cell recognition; however results showed that despite being genetically diverse, few predicted epitopes were found within the PGRS domain. The authors concluded that human T-cell recognition is not a significant factor driving sequence diversity within the PE_PGRS proteins (Copin *et al*. 2014). Similar analysis has not been performed for the PPE_MPTR proteins, and the role of the extensive genetic variation observed within the PPE_MPTR proteins is unclear, particularly whether genetic diversity within these genes contributes to the pathogenicity of *M. tuberculosis*. Whether or not the PPE_MPTR proteins are a source of antigenic variation is therefore yet unanswered.

To date no studies have systematically addressed the role of the PPE_MPTR proteins in host-pathogen interactions, and given the possible role of these proteins in the interaction with the immune system, they may be candidates for potential vaccine targets that should be further explored.

1.2 Problem Statement

Members of the PPE_MPTR family are known to be unique to pathogenic mycobacteria and are thought to play a role in host-pathogen interactions. Nevertheless this set of proteins is relatively poorly characterized, and the exact mechanisms and consequences of the host-pathogen interactions are not well understood. The presence and distribution of epitopes within this sub-family of *M. tuberculosis* proteins have also not been fully explored. The PPE_MPTR proteins are known to be highly variable, leading to the possible assumption that these proteins may play a role in antigenic variation seen in other pathogens but not (yet) in *M. tuberculosis*. Determining the mechanisms underlying *ppe_mptr* diversification will also aid in the

understanding of the role of these genes in the evolution of mycobacterial pathogenicity. A better understanding of the role of the PPE_MPTR proteins in modulating or evading host immune responses will aid in the development of novel interventions such as vaccines in order to favourably influence the outcome of infections with *M. tuberculosis*.

Conventional wet lab approaches used to ascertain host-pathogen interactions are often time-consuming and costly due to the large number of possible experiments needed. Computational methods offer the opportunity to use the large amount of existing data together with mathematical and statistical techniques to make accurate predictions about the interactions between *M. tuberculosis* and host components. This will allow researchers to gain a better understanding of host-pathogen interactions and will also inform the type and level of future experiments needed.

1.3 Hypotheses

Given the hyper-variable nature of the PPE_MPTR proteins and their possible role in modulating host-pathogen interactions, genetic diversity within PPE_MPTR proteins may differentially modulate human immune response.

Epitopes within the PPE_MPTR proteins may be possible subunit vaccine candidates for *M. tuberculosis*.

1.4 Aims and Objectives

1.4.1 Aim 1: *In silico* identification of human CD4+ T-cell epitopes in PPE_MPTR proteins

CD4+ T-cell epitopes induce an immune response when bound by HLA/MHC class II molecules presented on the cell surface for recognition by T-cells. The conventional approach to identifying T-cell epitopes involves synthesizing overlapping peptide fragments and assaying for binding to HLA class II molecules (Figure 1.2). However this method can become time consuming and costly, especially for large proteins such as certain members of the PPE_MPTR family. Computational methods have been developed which predict binding of peptides to HLA molecules using a variety of statistical algorithms based on known epitope data.

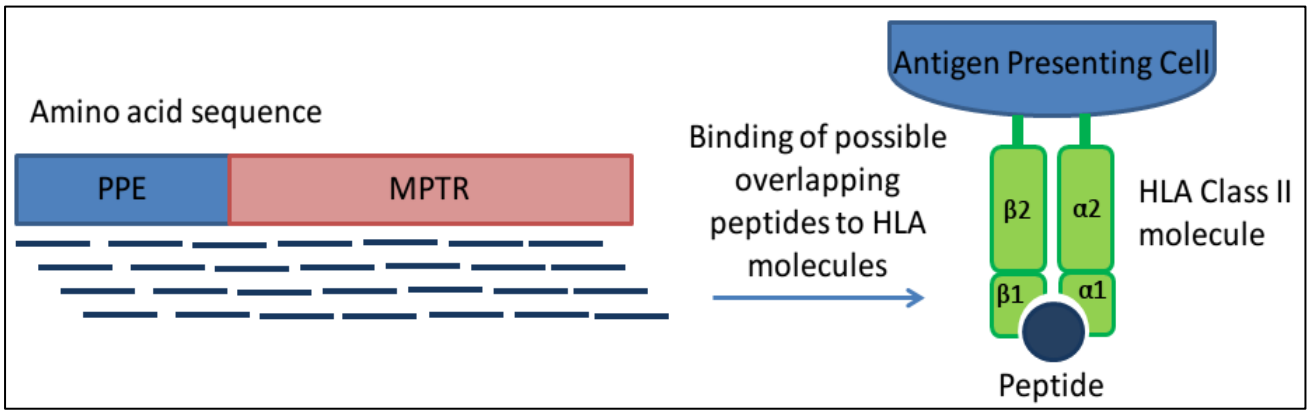


Figure 1.2: Methodology for identifying potential epitopes. Binding of overlapping peptides along the length of each PPE_MPTR protein with various HLA class II alleles is predicted using *in silico* approaches.

1.4.1.1 Objective 1: Determination of the optimal epitope prediction pipeline

There are currently more than 20 predictive tools used to predict binding of peptides to HLA class II molecules. Using a collection of known *M. tuberculosis* epitopes from the Immune Epitope Database (IEDB), an evaluation of the current open source HLA class II prediction tools will be performed. As a result, the optimal epitope prediction pipeline for *M. tuberculosis* proteins will be determined.

1.4.1.2 Objective 2: Prediction of PPE_MPTR epitopes

Prediction of binding of overlapping peptides within each of the PPE_MPTR to various HLA alleles will be performed using the pipeline determined in objective 1. Resulting predicted epitopes may be potential candidates for subunit vaccines, and provide a more targeted starting point for further analysis.

1.4.2 Aim 2: Characterisation of the level of genetic diversity within the PPE_MPTR family members

The number and type of mutational events such as single nucleotide polymorphisms (SNPs), insertions and/or deletions (indels) or large macro-mutations will be determined and compared for each PPE_MPTR protein. Publically available whole genome sequences (WGS) from strains belonging to various lineages will be used. Results from this analysis will provide insight into the mechanisms underlying PPE_MPTR diversification and may shed light on the role that these proteins play in mycobacterial pathogenicity.

1.4.3 Aim 3: Determine the impact of genetic variants on epitope prediction

Investigating the presence or lack of genetic variants in areas of predicted epitopes will help determine whether genetic diversity within PPE_MPTR T-cell epitopes differentially modulates human immune response as hypothesized, and therefore whether these proteins may represent a source of antigenic variation. These results will also be an important consideration in determining whether epitopes from the PPE_MPTR proteins are potential sub-unit vaccine candidates.

1.4.4 Aim 4: Identification of potential vaccine candidates

Predicted epitopes will be filtered using a reverse vaccinology (RV) approach in order to identify possible vaccine candidates.

1.5 Significance of Research

This research may provide novel insights into the function of the PPE_MPTR proteins, and whether they may play a role in host-pathogen interaction.

Possible vaccine candidates may be identified that may prompt further research and experimentation.

1.6 Scope and Limitations

This project is limited to the PPE_MPTR proteins of *M. tuberculosis* and to the prediction of MHC/HLA class II responses.

Results are based on purely computational approaches which provide the most likely answer given a set of input data. Experimental validation of predictions has not been performed. The results in this project therefore provide a starting point for further analysis and validation.

The *pe/ppe* genes are notoriously difficult to sequence as well as to align/map to a reference genome due to the high sequence variation and repeat regions found within these genes. The short reads of the current sequencing technology used can result in low confidence mapping over these regions and therefore in the resulting variants called. Assembled WGS data used within this project has been gathered from public databases and has not been checked for quality. It has been assumed that the assembly of each WGS is acceptable, and the resulting gene sequences within the PE/PPE regions are correct. Where possible, additional steps have been taken to ensure data is accurate, however no additional targeted sequencing has been performed to validate the genetic variation identified.

1.7 Chapter Overview

Chapter 2: **Literature review**
The use of immunoinformatics in the identification of vaccine candidates for *M. tuberculosis*

Chapter 3: **Evaluation of MHC class II epitope prediction tools**
Eight prediction tools have been evaluated using experimentally validated *M. tuberculosis* CD4+ T-cell epitopes.

Chapter 4: **MHC II epitope prediction**
This chapter presents the results from an *in silico* identification of MHC class II epitopes within the PPE_MPTR proteins. Various HLA class II alleles have been included.

Chapter 5: **Genetic diversity of PPE_MPTR proteins**
This chapter presents genetic diversity results from the analysis of 33 *M. tuberculosis* strains and 5 other mycobacterial strains. The correlation between areas of diversity with areas of predicted epitopes has been evaluated.

Chapter 6: **Identification of potential vaccine candidates and conclusion**
Using the results from Chapter 4 and Chapter 5 potential vaccine candidates have been identified using a Reverse Vaccinology approach. This includes the *in silico* prediction of the potential population coverage of each possible vaccine candidate within South Africa. Concluding remarks are given.

1.8 References

Abubakar, I. et al., 2013. Systematic review and meta-analysis of the current evidence on the duration of protection by bacillus Calmette–Guérin vaccination against tuberculosis. *Health Technol Assess*, 17(37), pp.1–372, v–vi.

Akhter, Y. et al., 2012. The PE/PPE multigene family codes for virulence factors and is a possible source of mycobacterial antigenic variation: Perhaps more? *Biochimie*, 94(1), pp.110–116.

Bansal, K. et al., 2010. Src homology 3-interacting domain of Rv1917c of Mycobacterium tuberculosis induces selective maturation of human dendritic cells by regulating PI3K-MAPK-NF-kappaB signaling and drives Th2 immune responses. *The Journal of biological chemistry*, 285(47), pp.36511–22.

Basu, S. et al., 2007. Execution of macrophage apoptosis by PE_PGRS33 of Mycobacterium tuberculosis is mediated by Toll-like receptor 2-dependent release of tumor necrosis factor-alpha. *The Journal of biological chemistry*, 282(2), pp.1039–50.

Bertholet, S. et al., 2010. A defined tuberculosis vaccine candidate boosts BCG and protects against multidrug-resistant Mycobacterium tuberculosis. *Science translational medicine*, 2(53), p.53ra74.

Bertholet, S. et al., 2008. Identification of human T cell antigens for the development of vaccines against Mycobacterium tuberculosis. *Journal of immunology (Baltimore, Md. : 1950)*, 181(11), pp.7948–57.

Billeskov, R. et al., 2012. The HyVac4 subunit vaccine efficiently boosts BCG-primed anti-mycobacterial protective immunity. *PloS one*, 7(6), p.e39909.

Borst, P., 1991. Molecular genetics of antigenic variation. *Immunology today*, 12(3), pp.A29–33.

Brennan, M.J. et al., 2001. Evidence that mycobacterial PE_PGRS proteins are cell surface constituents that influence interactions with other cells. *Infection and Immunity*, 69(12), pp.7326–7333.

Brodin, P. et al., 2010. High content phenotypic cell-based visual screen identifies Mycobacterium tuberculosis acyltrehalose-containing glycolipids involved in phagosome remodeling. *PLoS pathogens*, 6(9), p.e1001100.

Caruso, A.M. et al., 1999. Mice deficient in CD4 T cells have only transiently diminished levels of IFN-gamma, yet succumb to tuberculosis. *Journal of immunology (Baltimore, Md. : 1950)*, 162(9), pp.5407–16.

Choudhary, R.K. et al., 2004. Expression and characterization of Rv2430c, a novel immunodominant antigen of Mycobacterium tuberculosis. *Protein expression and purification*, 36(2), pp.249–53.

Cole, S.T. et al., 1998. Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. *Nature*, 393(6685), pp.537–44.

Comas, I. et al., 2010. Human T cell epitopes of Mycobacterium tuberculosis are evolutionarily hyperconserved. *Nature genetics*, 42(6), pp.498–503.

Copin, R. et al., 2014. Sequence diversity in the pe_pgrs genes of Mycobacterium tuberculosis is independent of human T cell recognition. *mBio*, 5(1), pp.e00960–13.

Day, C.L. et al., 2013. Induction and regulation of T-cell immunity by the novel tuberculosis vaccine M72/AS01 in South African adults. *American journal of respiratory and critical care medicine*, 188(4), pp.492–502.

Delany, I., Rappuoli, R. & De Gregorio, E., 2014. Vaccines for the 21st century. *EMBO molecular medicine*, 6(6), pp.708–20.

Dworkin, J. & Blaser, M.J., 1997. Molecular mechanisms of Campylobacter fetus surface layer protein

expression. *Molecular microbiology*, 26(3), pp.433–40.

Ernst, J.D., 2012. The immunological life cycle of tuberculosis. *Nature reviews. Immunology*, 12(8), pp.581–91.

Flynn, J.L. et al., 1992. Major histocompatibility complex class I-restricted T cells are required for resistance to Mycobacterium tuberculosis infection. *Proceedings of the National Academy of Sciences of the United States of America*, 89(24), pp.12013–7.

Gey van Pittius, N.C. et al., 2006. *Evolution and expansion of the Mycobacterium tuberculosis PE and PPE multigene families and their association with the duplication of the ESAT-6 (esx) gene cluster regions.*,

Hagblom, P. et al., 1985. Intragenic recombination leads to pilus antigenic variation in Neisseria gonorrhoeae. *Nature*, 315(6015), pp.156–8.

Ireton, G.C. et al., 2010. Identification of Mycobacterium tuberculosis antigens of high serodiagnostic value. *Clinical and vaccine immunology : CVI*, 17(10), pp.1539–47.

Karboul, A. et al., 2008. Frequent homologous recombination events in Mycobacterium tuberculosis PE/PPE multigene families: potential role in antigenic variability. *Journal of bacteriology*, 190(23), pp.7838–46.

Keating, G.M. & Noble, S., 2003. Recombinant hepatitis B vaccine (Engerix-B): a review of its immunogenicity and protective efficacy against hepatitis B. *Drugs*, 63(10), pp.1021–51.

Koomey, M., 1997. Bacterial pathogenesis: a variation on variation in Lyme disease. *Current biology : CB*, 7(9), pp.R538–40.

Kozakiewicz, L. et al., 2013. The role of B cells and humoral immunity in Mycobacterium tuberculosis infection. *Advances in experimental medicine and biology*, 783, pp.225–50.

Kwan, C.K. & Ernst, J.D., 2011. HIV and tuberculosis: a deadly human syndemic. *Clinical microbiology reviews*, 24(2), pp.351–76.

Lew, J.M. et al., 2011. TubercuList--10 years after. *Tuberculosis (Edinburgh, Scotland)*, 91(1), pp.1–7.

Loxton, A., Hondalus, M. & Sampson, S., 2016. TB vaccine assessment. In P. B. F. Anthony J. Hickey, Amit Misra, ed. *Delivery Systems for Tuberculosis Prevention and Treatment*. Wiley, pp. pp91–110.

Luca, S. & Mihaescu, T., 2013. History of BCG Vaccine. *Mædica*, 8(1), pp.53–8.

McEvoy, C.R.E. et al., 2009. Evidence for a rapid rate of molecular evolution at the hypervariable and immunogenic Mycobacterium tuberculosis PPE38 gene region. *BMC evolutionary biology*, 9, p.237.

McShane, H. et al., 2005. Boosting BCG with MVA85A: the first candidate subunit vaccine for tuberculosis in clinical trials. *Tuberculosis (Edinburgh, Scotland)*, 85(1-2), pp.47–52.

Murphy, K., 2011. *Janeway's Immunobiology* 8th ed., Garland Science.

Nascimento, I.P. & Leite, L.C.C., 2012. Recombinant vaccines and the development of new vaccine strategies. *Brazilian journal of medical and biological research = Revista brasileira de pesquisas médicas e biológicas / Sociedade Brasileira de Biofísica ... [et al.]*, 45(12), pp.1102–11.

Nassif, X. et al., 1993. Antigenic variation of pilin regulates adhesion of Neisseria meningitidis to human epithelial cells. *Molecular microbiology*, 8(4), pp.719–25.

Philips, J.A. & Ernst, J.D., 2012. Tuberculosis pathogenesis and immunity. *Annual review of pathology*, 7, pp.353–84.

Sampson, S.L. et al., 2001. Expression, characterization and subcellular localization of the Mycobacterium

tuberculosis PPE gene Rv1917c. *Tuberculosis (Edinburgh, Scotland)*, 81(5-6), pp.305–317.

Sampson, S.L., 2011. Mycobacterial PE/PPE proteins at the host-pathogen interface. *Clinical and Developmental Immunology*, 2011(Figure 1).

Sani, M. et al., 2010. Direct visualization by cryo-EM of the mycobacterial capsular layer: a labile structure containing ESX-1-secreted proteins. *PLoS pathogens*, 6(3), p.e1000794.

Song, L. et al., 2015. An avian influenza A (H7N9) virus vaccine candidate based on the fusion protein of hemagglutinin globular head and Salmonella typhimurium flagellin. *BMC biotechnology*, 15, p.79.

Stewart, G.R. et al., 2005. Mycobacterial mutants with defective control of phagosomal acidification. *PLoS pathogens*, 1(3), pp.269–78.

WHO, 2015. *Global tuberculosis report 2015*, World Health Organization.

van der Woude, M.W. & Baumler, A.J., 2004. Phase and Antigenic Variation in Bacteria. *Clinical Microbiology Reviews*, 17(3), pp.581–611.

Chapter 2: Literature Review:

The use of immunoinformatics in the identification of vaccine candidates for *Mycobacterium tuberculosis*

2.1 Introduction

2.1.1 The search for a new vaccine against tuberculosis

The need for a new vaccine against Tuberculosis (TB) has been well documented, with the current Bacillus Calmette-Guérin (BCG) vaccine, developed more than 80 years ago, showing varying levels of efficacy (WHO 2015). Several promising vaccines are currently in various phases of clinical trials (Chapter 1, Table 1.1), which include subunit or peptide based vaccines which consist of selected mycobacterial antigens (Loxton *et al*. 2016). Identifying the optimal composition of protective antigens is an essential step towards developing new peptide based vaccines (Vivona *et al*. 2008). Protective immunity against *Mycobacterium tuberculosis* is induced by stimulating antigen specific T-cells which recognise peptide antigens presented on human leukocyte antigen (HLA) molecules (Figure 2.1). Bacterial peptides that are recognised by the host immune system in this way are called T-cell epitopes, and identifying epitopes that are capable of binding to HLA molecules and eliciting T-cell responses forms part of the development of subunit vaccines for *M. tuberculosis*.

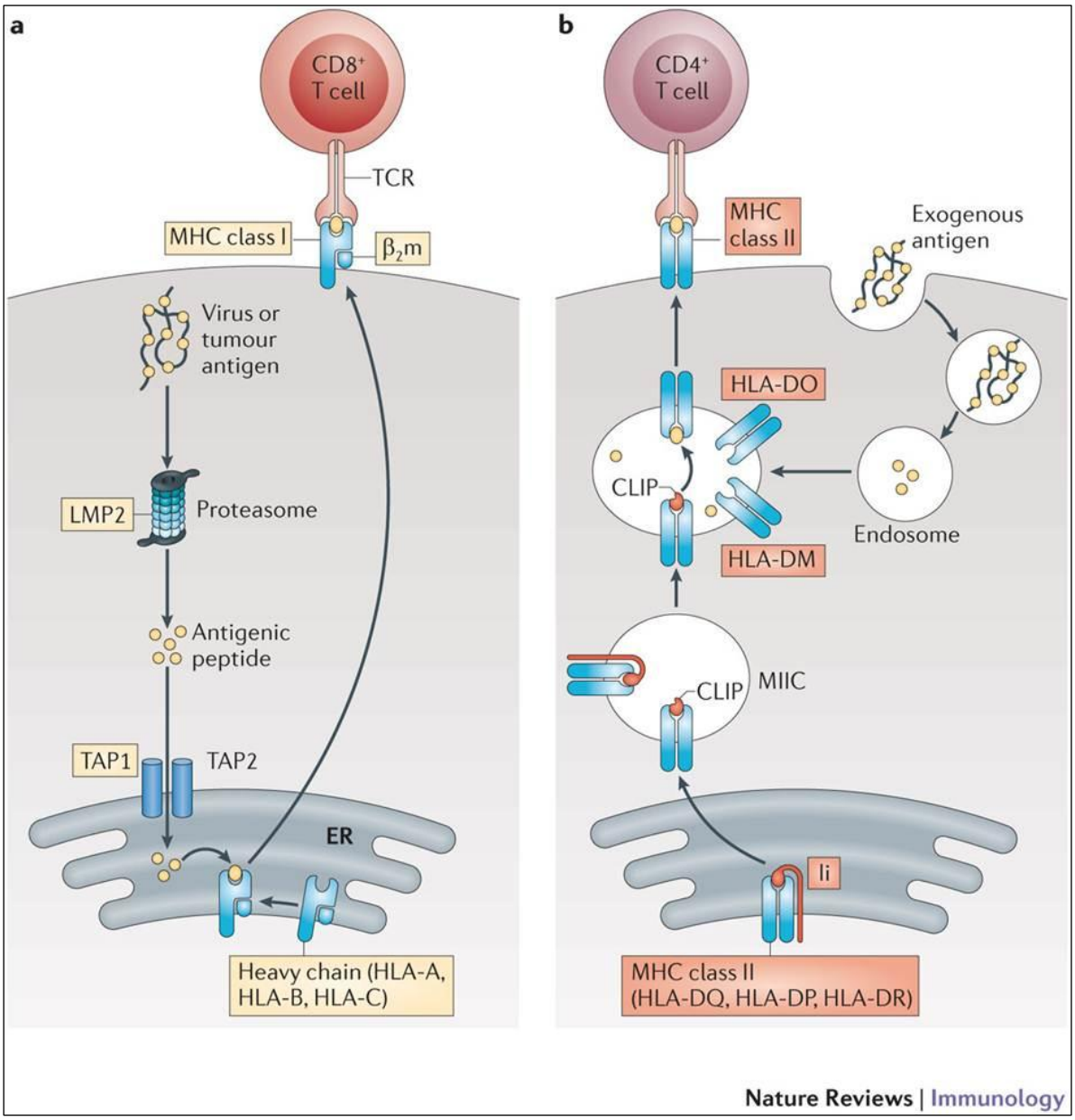


Figure 2.1: Antigen Presentation Pathways. a) Intracellular antigens are degraded by the immunoproteasome and transported to the endoplasmic reticulum where they are loaded into HLA class I molecules and presented to CD8+ T-cells. b) Antigens from extracellular sources, which include bacterial pathogens are degraded by endolysosomal enzymes, and loaded into HLA class II molecules and presented to CD4+ T-cells (Kobayashi & van den Elsen 2012).

Experimental techniques used to identify potential epitopes can often be time consuming and expensive; however *in silico* approaches such as those used in immunoinformatics provide an alternative option.

2.1.2 Immunoinformatics

A large amount of immunological data has accumulated due to clinical research, epidemiological investigations and developments in sequencing of both human and pathogen genomes. This has led to new discoveries in immune function and disease pathogenesis (Tomar & De 2014). This growing amount of immunological data as well as the *in silico* techniques used to analyse this data has led to the emergence of a specific branch of bioinformatics known as immunoinformatics. Immunoinformatics can be described as the interface between computer science and experimental immunology (Tomar & De 2014). Having access to this wealth of data along with the tools available to analyse it has made searching for vaccine targets a much more efficient process. The use of immunoinformatic techniques in the search for potential vaccine candidates is termed reverse vaccinology (RV). RV workflows aim to narrow down the potential list of antigens to be included in a vaccine which can then be further tested and validated in an experimental setting, thereby lowering the time and cost associated with laboratory analysis. The applicability and usefulness of RV techniques was first demonstrated by Pizza *et al*. (2000) in the identification of vaccine candidates against Serogroup B Meningococcus (MenB). In this seminal study, whole genome sequence (WGS) data was used to predict novel surface exposed or exported proteins, followed by comparative genomics in order to test the suitability of the candidate antigens to induce protection against various clinical MenB strains. Since then RV techniques have been applied to various other pathogens including *M. tuberculosis*.

The accumulation of the large amount of genomic and experimentally generated immunological data has necessitated the need for immunological databases. These databases are needed to not only store and categorise data but also enable sharing of the data between researchers. Prediction methods used in RV workflows also require large amounts of training data in order to make the algorithms as accurate as possible. Immunoinformatics therefore acts as an intersection between immunological data and the computational techniques and algorithms used to make predictions (Vivona *et al*. 2008).

A traditional RV approach includes the use of comparative genomics; epitope prediction; analysis of protein properties; and the prediction of protective coverage to identify immunogenic antigens (Figure 2.2). Certain workflows within a traditional RV approach may be more or less applicable depending on the pathogen being studied, and actual pipelines used should be individually tailored. In certain cases, selection criteria (i.e. by type or function of protein) may be used in order to first narrow down the list of possible antigenic targets being fed into a RV workflow. The selection criteria should reflect the different mechanisms contributing to the pathogenesis of the pathogen. However, in certain cases, a bias may be introduced where the actual selection criteria is a reflection of the current knowledge (or lack thereof) of the biology of the pathogen in question. In other cases, experimental limitations may result in selection bias focusing on proteins that can be studied in a wet lab setting. In these cases, a whole genome approach may overcome any selection bias introduced.

Figure 2.2: Traditional Reverse Vaccinology (RV) Workflow. A traditional RV workflow used to identify possible vaccine targets generally includes comparative genomics, epitope prediction, analysis of protein properties and prediction of protective coverage (Brusic & Petrovsky 2005; Vivona *et al*. 2008; Movahedi & Hampson 2008; Mora *et al*. 2006; Serruto *et al*. 2009; Davies & Flower 2007; Bambini & Rappuoli 2009; Rapin *et al*. 2010; Tomar & De 2010).

2.1.3 Adaption of RV workflows to *M. tuberculosis*

M. tuberculosis has various unique characteristics, not only in structure but in the interaction of the bacteria with the human host. For example, given the importance of T-cells in initiating and regulating the adaptive immune response against *M. tuberculosis*, and the relatively unclear role of B-cells and antibodies, TB researchers have primarily focused on the prediction of T-cell rather than B-cell epitopes. In addition, when choosing selection criteria, many RV workflows limit searching for possible vaccine candidates to surface exposed/secreted proteins which may be involved in adhesion, invasion, secretion and signalling host responses (Vivona *et al*. 2008). While these may be important candidates, limiting the search to these proteins may not be appropriate for *M. tuberculosis.* Immunity against *M. tuberculosis* is primarily induced via stimulation of CD4+ T-cells, which recognise peptides derived from the degradation of bacterial proteins which have been internalized by macrophages and presented by HLA class II molecules. In addition, apoptosis of the infected cells can result in bacterial fragments taken up by various antigen presenting cells which initiates the HLA class I presentation pathway and CD8+ T-cell stimulation (Figure 2.1). Literature has shown that BCG vaccination (which primarily induces CD4+ immunity) can be enhanced by promoting antigen translocation into the cytoplasm and stimulating CD8+ T-cells, therefore enhancing overall T-cell mediated immunity and superior protection against *M. tuberculosis* (Grode *et al*. 2005). Subunit vaccines combining peptides capable of inducing both CD4+ as well as CD8+ immunity may therefore be most effective against *M. tuberculosis*. For this reason, limiting RV searches to surface exposed/secreted proteins may underestimate the range of possible targets.

RV workflows used in the analysis of protein properties may also not be as applicable when considering *M. tuberculosis*. For example, bacterial lipoproteins, may represent a class of possible vaccine candidates for many pathogens given their role in cell signalling and substrate binding, for which *in silico* tools such as LipPred have been developed to identify (Taylor *et al*. 2006). These tools are usually developed specifically for either Gram-positive or Gram-negative bacteria. However, while *M. tuberculosis* belongs to the group of Gram-positive bacteria, the lipid-rich outer layer of *M. tuberculosis* is comparable to the outer membrane of Gram-negative bacteria (Rezwan *et al*. 2007). These *in silico* tools may therefore be less accurate when used to identify lipoproteins for *M. tuberculosis*.

The level of genetic diversity for different pathogens will determine the importance of comparative genomics within the RV workflow. In 1998, the complete genome sequence of a well characterised strain of *M. tuberculosis*, H37Rv, was first completed (Cole *et al*. 1998). Since then, a large number of strains from different lineages have been sequenced allowing for comparative genomic analysis between strains. These investigations have helped explain the evolution of *M. tuberculosis* and identification of essential genes and others contributing to pathogenicity. Conserved versus genetically diverse regions within the genome have also been identified, making comparative genomics an essential step within the RV workflow as potential vaccine candidates should be effective against various clinical strains of *M. tuberculosis*. This is particularly relevant when considering the *ppe_mptr* genes as each protein contains both a conserved region (PPE) and a highly variable region (MPTR). Genetic variation of the *ppe_mptr* genes across various strains is investigated in Chapter 5 and the results used to filter for possible vaccine candidates in Chapter 6.

In addition to WGS data, large amounts of transcriptomic and proteomic data has been generated. Combining this omics information with data generated through experimental laboratory analysis has given new insights into the function of *M. tuberculosis* proteins, and their role within host-pathogen interactions. However, many encoded proteins within the *M. tuberculosis* genome still have an unknown or hypothetical function, some of which may play a possible role in human immune response. Immunoinformatic techniques offer researchers the opportunity to use known data to model various immunological processes, leading to a better understanding of the systems and pathways that may be involved in host-pathogen interactions and therefore aiding in the search for new vaccine candidates, even within proteins with an unknown or hypothetical function. In this way, selection bias imposed by a limited knowledge of the mechanisms of pathogenesis for *M. tuberculosis* may be avoided through RV.

RV workflows that include comparative genomics, epitope prediction and analysis of protein properties have been applied to *M. tuberculosis*. For example, the MycobacRVWebserver is a database containing the results of a RV workflow applied to 23 mycobacterial strains (including 13 *M. tuberculosis* strains, and 10 other pathogenic mycobacterial proteomes) (Chaudhuri *et al*. 2014). These results are available to TB researchers. In an alternative study, a RV workflow was applied to the *M. tuberculosis* H37Rv strain using the New Enhanced Reverse Vaccinology Environment (NERVE) software (Monterrubio-López *et al*. 2015), which has been developed to provide researchers with multiple well-known algorithms for protein analysis and comparison (Vivona *et al*. 2006).

In addition to the above, there are a substantial number of other immunological databases and prediction methods available to TB researchers which can be used in the search for possible vaccine targets. Databases containing T-cell epitope information and tools related to predicting T-cell epitopes and their expected protective coverage have been focused on here. This chapter provides a review of these resources including examples where *in silico* immunological techniques have already assisted in identifying vaccine candidates for *M. tuberculosis*.

2.2 Immunological Databases

2.2.1 T-Cell epitope databases

Immunological databases contain a wide variety of data sources including data from laboratory experiments performed to elucidate mechanistic aspects of an immune system and/or response to various infections (Vita *et al*. 2014). A large proportion of these experiments include identifying T cell epitopes. This typically involves synthesizing overlapping peptides from a pathogen protein and performing assays to test for binding to HLA alleles and/or cellular immune responses of the host to exposure of the antigen peptide. Table 2.1 shows current immunological databases containing T-cell epitope data, which are actively being updated and contain information relating to mycobacteria. Other databases, such as InnateDB which focuses on pathways and interactions involved in the innate immune response rather than the adaptive immune response are also available (Breuer *et al*. 2013).

Table 2.1: T-Cell Epitope Databases. Databases that are currently accessible, actively being updated, and contain information relating to mycobacteria have been included.

Database	Reference
IEDB (Immune Epitope Database) Contains experimental data characterising epitopes (B-cell and T-cell) involved in infectious diseases (in human and other animal species), allergy, autoimmunity and transplant. Data is curated from published literature and submitted by NIH funded epitope discovery efforts.	(Vita *et al*. 2014) www.iedb.org
Syfpeithi Contains data relating to peptide anchor motifs and MHC ligands of humans and other animal species, including peptide sequences, anchor positions, MHC specificity, source organisms and publication references. Data is curated from published literature.	(Rammensee *et al*. 1999) www.syfpeithi.de
AntiJen (Previously known as JenPep) Contains quantitative binding data for peptides binding to MHC Ligand, TCR-MHC Complexes, B-cell and T-cell epitopes, TAP and immunological protein-protein interactions. Additionally, AntiJen integrates this cellular data with kinetic, thermodynamic and functional information. Data is curated from published experimentally determined experiments.	(Blythe *et al*. 2002; McSparron *et al*. 2003; Toseland *et al*. 2005) http://www.ddg-pharmfac.net/antijen/AntiJen/antijenhomepage.htm
MHCBN Contains detailed information about MHC binding and non-binding peptides and T-cell epitopes, peptides interacting with TAP and MHC linked autoimmune diseases.	(Lata *et al*. 2009) http://www.imtech.res.in/raghava/mhcbn/

The collective knowledge of known *M. tuberculosis* epitopes contained within databases such as the IEDB not only aids in vaccine development but can also identify gaps and suggest potential areas for further research and discovery. In 2007, a comprehensive analysis of *M. tuberculosis* data contained within the IEDB was performed, with results showing that at the time epitopes were reported in only 7% of all open reading frames, with the top 30 studied protein antigens containing approximately 65% of reported epitopes (Blythe *et al*. 2007). It is evident from this study that the majority of the *M. tuberculosis* proteome has been under explored in the context of epitope identification. The majority of epitopes were found in proteins associated with virulence, detoxification and adaption; however epitope density of mycobacterial proteins may not be indicative of the true distribution of epitopes within the proteome but rather due to experimental bias. RV workflows may therefore be helpful in unbiased identification of epitopes across the proteome without the limitations imposed by experimental analysis such as difficulty of protein isolation. This study also highlighted an important shortfall of immunological databases in that information relating to the specific strain from which the epitope was derived is not recorded, which is an important consideration for a pathogen such as *M. tuberculosis* which contains polymorphic areas within its genome.

2.2.2 Human immune cell databases

Databases containing genomic information of human immune cells, predominantly focused on HLA molecules, are also available to immunological researchers (Table 2.2). These databases aim to capture the different alleles reported across various populations resulting from the polymorphic nature of the human genome. Both the genetic information of the bacteria and the human host are extremely important when studying host-pathogen interactions, particularly when trying to determine the protective coverage of a vaccine target (Section 2.5).

Table 2.2: Databases containing human immune data. Human immune cell databases available to TB researches with corresponding URLs are given.

Database	Reference
dbMHC Contains DNA and clinical data related to HLA molecules. Includes an interactive alignment viewer for HLA and related genes, an MHC microsatellite database, a sequence interpretation site for sequencing based typing and a primer/probe database.	(Helmberg *et al*. 2004) http://www.ncbi.nlm.nih.gov/projects/gv/mhc/
IMGT (International ImMunoGeneTics information system) Contains data relating to immunoglobulins/antibodies, T-cell receptors, MHC and HLA molecules, and related proteins of the immune system. Genome sequences as well as structural information is included.	(Lefranc 2008) www.imgt.org
AFND (Allele Frequency Net Database) Provides allele frequencies in different populations across the world from polymorphic areas in the human genome which are involved in immune functions including HLA, killer cell immunoglobulin-like receptors (KIR), cytokines and major histocompatibility complex class I chain-related genes (MIC).	(González-Galarza *et al*. 2015) www.allelefrequencies.net

2.2.3 Mycobacterium specific databases

Databases collecting and storing mycobacteria specific data include Tuberculist (Lew *et al*. 2011) and TBDB (Galagan *et al*. 2010). These databases provide a valuable source of WGS data which can be used to perform comparative genomic analysis of various *M. tuberculosis* strains. Expression data which can provide a starting point for selection criteria or alternatively be used as a potential source of validation following the identification of potential vaccine candidates is also available. TBDB also provides a complete list of epitopes for the clinical strain H37Rv stored within the IEDB.

2.2.4 Data repositories for machine learning

Predictive tools used in reverse vaccinology workflows rely on known experimental data to train machine learning algorithms (section 2.3.1). The accuracy of prediction tools is therefore dependent on the amount and quality of this data. In addition to the databases shown in Table 2.1, data repositories such as the Dana-Farber Repository for Machine Learning in Immunology (DFRMLI) have been created. DFRMLI aims to bridge the gap between immunologists and computer scientists by providing pre-processed and scaled immunological datasets which can be used in machine learning applications (Zhang *et al*. 2011).

2.3 T-Cell Epitope Prediction

Traditional methods of epitope identification involve the synthesis and testing of overlapping peptides that span the entire protein or target antigen using experimental assays. However this systematic mapping approach can become costly and time consuming, especially for large proteins within the *M. tuberculosis* proteome and the large repertoire of HLA molecules that exists. This has necessitated the need to develop *in silico* epitope prediction methods.

HLA binding is required for T-cell recognition, and therefore *in silico* T-cell epitope prediction methods rely heavily on predicting the binding affinity of short peptide sequences to various HLA class I and class II molecules. Various steps are involved during antigen presentation, of which binding to an HLA molecule is just one. The majority of epitope prediction tools are therefore indirect epitope prediction methods since it is assumed that a peptide able to bind to an HLA molecule will illicit an immune response. *In silico* methods which model and/or predict other parts of the HLA class I antigen presentation pathway such as proteasomal cleavage and the transfer-associated protein (TAP) are also available to researchers and are discussed below.

There are currently over 30 different HLA class I and HLA class II prediction tools openly available. These tools are based on various prediction algorithms and methods. The majority of these methods involve defining peptide-HLA binding specificities based on large datasets of known HLA binding peptides. These datasets are used to train predictive algorithms.

2.3.1 Prediction algorithms

Various prediction algorithms have been used to predict T-cell epitopes (Table 2.3). These methods are grouped into either direct prediction methods which are based on the sequence and/or structural analysis of T-cell epitopes such as binding motifs; or indirect methods which are based on statistical learning theory.

Table 2.3: Prediction Algorithms. Both direct and indirect prediction algorithms are used within various epitope prediction tools.

Algorithm	Description
Direct Methods: Sequence/Motif and structural based.	These algorithms involve searching for known HLA binding amino acid motifs called anchor motifs. Direct methods include using position specific scoring matrices (PSSM) or quantitative matrices that take into account the position of each amino acid along the length of a peptide and assigns a value based on the frequency of occurrence in known T-cell epitopes. Methods using predicted structural information such as presence of hydrophilic and hydrophobic regions from sequence information also fall into this category. Proteochemometrics-based approaches have also been applied to HLA class II binding prediction. These approaches use quantitative structure-activity relationship (QSAR) and partial least squares regression analysis (PLS) which relates quantitative properties of ligands to their activity. These properties can include volume, hydrophobicity, and polarizability for each of the amino acid residues. Specificity-determining residues (SDR) have been used to describe amino acid residues that are responsible for specific interactions between a pair of proteins or in the case of HLA binding prediction, a protein and a peptide. In this case, developers have used crystal structures of peptide and HLA complexes.
Indirect Methods: Machine learning methods	Indirect Methods include using Artificial Neural Networks (ANN), Hidden Markov Models (HMM), and Support Vector Machines (SVM) (Vivona *et al*. 2008). ANN's are trained to learn features of appropriate patterns from known data of peptide and HLA binding, and subsequently recognise similar patterns in new data with unknown binding ability. HMM's are statistical models which use a finite set of states, each with an associated probability distribution to determine the most likely state of the process based on the given input data. SVM's are trained to recognise patterns present in known data by mapping the data into a high-dimensional feature space where every coordinate corresponds to a particular feature within the data.

2.3.2 Data quantity and quality

Given that the majority of these prediction methods make use of known epitope data to train the predictive algorithm, the accuracy of each model is dependent on the quantity and quality of training data used, further highlighting the need for the storage and access to immunological data.

There exists a high peptide to HLA molecule specificity and therefore a sufficient amount of binding affinity data for each HLA allele is needed to accurately make predictions per allele. Based on this, the accuracy, sensitivity, and specificity of each prediction method differ per HLA allele and is reflective of the amount of HLA data included in the training set used. This has been demonstrated for both HLA class I and class II alleles during a comparative evaluation of the publically available prediction servers (Lin, Ray, *et al*. 2008; Lin, Zhang, *et al*. 2008). HLA molecules are extremely polymorphic, with over 15,635 alleles reported to date on the IGMT/HLA database (accessed 1 November 2016). This makes including data for each HLA allele in a training set and therefore prediction method almost impossible. Pan-specific methods have been developed in response to this need. Pan-specific methods extend prediction to a wider set of HLA molecules (for which little binding data is available), by extrapolating from the binding specificities of the molecules for which

sufficient binding data is available. This is done using similarities in the binding cores of the HLA molecules (Karosiene *et al*. 2013). However the accuracy of these pan-specific methods is generally lower than for other prediction methods (Lin, Zhang, *et al*. 2008).

In addition to using already trained algorithms, researchers are also able to customize their own predictions using their own data sources. EPIMHC is a database of T-cell epitopes and HLA binding peptides and allows a user to choose which data to use when constructing motif profiles (Reche *et al*. 2005). In this way, *M. tuberculosis* researchers are able to train a motif based algorithm using only known *M. tuberculosis* epitopes, while excluding other pathogens. This may improve the accuracy when investigating unknown *M. tuberculosis* proteins.

2.3.3 Epitope prediction tools

Table 2.4 includes a list of the currently available HLA class I and class II prediction tools available.

In general, the accuracy of HLA class I tools is higher than that reported for HLA class II prediction tools (Nielsen *et al*. 2010). While both HLA class I and class II molecules are heterodimers, with binding grooves consisting of a B-sheet and two alpha-helices, structural differences between the two types of HLA molecules affect the accuracy of the predictions. HLA class I molecules have a closed binding groove that binds an entire peptide (between 8-11 amino acids long); while HLA class II molecules bind peptides between 14-18 amino acids long, but have an open binding groove that accommodates a 9-mer binding core with the rest of the peptide binding on either side of the binding groove (Nielsen *et al*. 2010). There is a high binding specificity between the 9 mer binding core and the binding groove, while binding of the ends of the peptide to the edges of the binding groove is less specific. For HLA class I prediction tools, alignment free methods can be easily used since HLA class I peptides are generally of equal length and binding motifs have been well characterised, whereas HLA class II epitopes van vary greatly in length making alignment a crucial part of estimating binding motifs and predicted binding (Nielsen *et al*. 2010). This additional complexity has resulted in lower reported accuracy. As a result it has been suggested that HLA class I predictors should be used for prediction of positive binders while HLA class II predictors should be used for the elimination of obvious negatives or non-binders (Pappalardo *et al*. 2009). Given the objective of finding potential epitopes which will be validated in an experimental setting, researchers employing RV workflows often need to choose between decreasing the number of potential epitopes (i.e. increasing specificity) versus missing true epitopes (i.e. increasing sensitivity), particularly for HLA class II tools with a lower reported accuracy. Therefore eliminating obvious negatives is appropriate in the context of finding potential vaccine candidates (except in the case of extreme resource limitations). This is further discussed in Chapter 3.

Consensus approaches have been developed that use an average or median prediction value from various prediction tools. Evaluations have been performed comparing the accuracy of the various tools with one another (Lin, Ray, *et al*. 2008; Lin, Zhang, *et al*. 2008), as well as comparing the accuracy with consensus approaches (Wang *et al*. 2008; Wang *et al*. 2010; Moutaftsi *et al*. 2006). In general, consensus approaches have shown to be superior to individual approaches. Evaluation of HLA class II epitope prediction tools is further discussed in Chapter 3.

Table 2.4: T-cell Epitope Prediction Tools. A list of various prediction tools for both MHC class I and class II molecules with their corresponding URLs have been given.

Tool	Algorithm	Reference	URL
HLA Class I			
CTLPred	Quantitative matrices, ANN, SVM, Consensus	(Bhasin & Raghava 2004a)	http://www.imtech.res.in/raghava/ctlpred/
BIMAS	Matrix	(Parker *et al*. 1994)	https://www-bimas.cit.nih.gov/molbio/hla_bind/
IEDB	ANN, ARB, SMM	(Nielsen *et al*. 2003; Bui *et al*. 2005; Peters & Sette 2005)	http://tools.iedb.org/mhci/
MAPPP	Matrix	(Hakenberg *et al*. 2003)	http://www.mpiib-berlin.mpg.de/MAPPP/information.html
MHC BPS	SVM	(Cui *et al*. 2006)	
MHC-I	Structure-based model	(Jojic *et al*. 2006)	http://boson.research.microsoft.com/hlabinding/hlabinding.aspx
MHCPred	PLS	(Guan *et al*. 2006)	http://www.ddg-pharmfac.net/mhcpred/MHCPred/
MULTIPRED	ANN, HMM, SVM	(Zhang *et al*. 2005)	http://cvc.dfci.harvard.edu/multipred2/
NetMHC	ANN, Matrix	(Lundegaard *et al*. 2008)	http://www.cbs.dtu.dk/services/NetMHC/
nHLAPred	ANN and Matrix	(Bhasin & Raghava 2007)	http://www.imtech.res.in/raghava/nhlapred/
PepDist	Distance function	(Hertz & Yanover 2006)	http://www.mybiosoftware.com/pepdist-1-0-protein-peptide-binding-prediction.html
ProPred1	Matrix	(Singh & Raghava 2003)	http://www.imtech.res.in/raghava/propred1/
Rankpep	Matrix	(Reche *et al*. 2002)	http://imed.med.ucm.es/Tools/rankpep.html
SMM	Matrix	(Peters *et al*. 2003)	https://zlab.bu.edu/SMM/
SVMHC	SVM	(Dönnes & Kohlbacher 2006)	https://abi.inf.uni-tuebingen.de/Services/SVMHC
PeptideCheck	Matrix	(DeLuca *et al*. 2007)	http://www.peptidecheck.org/
HLA Class II			
EpiTop	QSAR	(Dimitrov *et al*. 2010)	http://www.pharmfac.net/EpiTOP/
Predivac	SDR	(Oyarzún *et al*. 2013)	http://predivac.biosci.uq.edu.au/
NetMHCII **NetMHCIIpan**	Matrix ANN	(Lundegaard *et al*. 2008; Karosiene *et al*. 2013)	http://www.cbs.dtu.dk/services/NetMHCII/ http://www.cbs.dtu.dk/services/NetMHCIIpan/
IEDB_ARB	Matrix	(Bui *et al*. 2005)	http://tools.iedb.org/mhcii/
NN-align	ANN	(Nielsen & Lund 2009)	http://tools.iedb.org/mhcii/
SMM-align	Matrix	(Nielsen *et al*.	http://tools.iedb.org/mhcii/

Tool	Algorithm	Reference	URL
		2007)	
IEDB Comblib	Matrix		http://tools.iedb.org/mhcii/
Sturniolo **Proped** **Syfpeithi**	Matrix	(Singh & Raghava 2001)	http://www.imtech.res.in/raghava/propred/
IEDB Consensus	Consensus	(Sidney *et al*. 2008)	http://tools.iedb.org/mhcii/
MULTIPRED	ANN, HMM, SVM	(Zhang *et al*. 2005)	http://cvc.dfci.harvard.edu/multipred2/
HLA-DR4Pred	ANN, SVM	(Bhasin & Raghava 2004b)	http://www.imtech.res.in/raghava/hladr4pred/
TEPITOPE **TEPITOPE-pan**	Matrix	(Bian & Hammer 2004; Zhang *et al*. 2012)	http://datamining-iip.fudan.edu.cn/service/TEPITOPEpan/TEPITOPEpan.html
ProPred	Matrix	(Singh & Raghava 2001)	http://www.imtech.res.in/raghava/propred/

2.3.4 HLA class I antigen presentation pathway tools

Direct epitope prediction tools which take into account the full HLA class I antigen presentation pathway (in addition to HLA class I binding) have been developed. This includes the prediction of proteasomal cleavage and TAP transport which are essential components of HLA class I antigen processing and presentation. TAP is responsible for delivering cytosolic peptides into the endoplasmic reticulum before loading them into the binding groove of an HLA class I molecule (Figure 2.1), and therefore TAP binding preferences can significantly impact T-cell epitope selection.

PRED^{TAP} (Zhang *et al*. 2006) and SVMTAP (Daniel *et al*. 1998) have been developed to predict binding of a peptide to human TAP. NetCTL (Larsen *et al*. 2005) ,WAPP (Dönnes & Kohlbacher 2005), EpiJen (Doytchinova *et al*. 2006) and MAPP (Hakenberg *et al*. 2003) are methods that integrate predictions of HLA class I binding, TAP transport and C-terminal proteasomal cleavage for an overall prediction of CTL epitopes.

2.4 Protective Coverage of Vaccine Candidates

2.4.1 HLA alleles

The genes encoding HLA molecules are extremely polymorphic, with a large number of HLA alleles reported to date. Genetic variations within the peptide binding groove of HLA molecules determine the type of peptide with which it is able to bind. There is therefore a high peptide to HLA molecule specificity, where a specific epitope may only bind to one, or in the case of promiscuous epitopes, a few HLA alleles. Currently 15,635 HLA and related alleles have been reported and are included in the IMGT/HLA database (Robinson *et al*. 2014) (accessed 1 November 2016). An individual person carries a limited number of alleles in their genome out of the thousands that are present in the population. Good subunit vaccine cocktails should therefore include a range of peptides that are able to bind to a wide range of HLA molecules in order to provide sufficient population coverage. Promiscuous epitopes which are able to bind to more than one HLA molecule may therefore be of particular importance in vaccine design, as they improve the potential population coverage when included in a vaccine cocktail. Certain HLA molecules show similar binding preferences and consequently HLA supertypes have been defined (Greenbaum *et al*. 2011). Given the similarity in the peptide binding groove within a particular supertype family, an epitope able to bind to one HLA molecule within a family should be able to bind to all of the HLA molecules within that family. As a result, an epitope-based vaccine can be tailored to be representative of HLA supertype families rather than the extensive repertoire of HLA molecules in order to increase potential population coverage.

The frequency of HLA alleles have been shown to exhibit an inter-ancestry distribution, with certain alleles more prevalent in certain ethnic populations than others. It has been suggested that 90% population coverage of several ethnic groups is possible by targeting 11 HLA molecules, however, in order to reach 90% coverage for African and Asian populations, an additional 4 or more alleles are required (Longmate *et al*. 2001). In South Africa particularly, the frequency of HLA class I and class II alleles differs significantly between Caucasian and Black populations (Paximadis *et al*. 2012). Black populations also show a broader spectrum of alleles compared to the single allele dominance seen in Caucasian populations (Paximadis *et al*. 2012). In a recent review, Tshabalala *et al*. (2015), investigated HLA diversity in Southern African populations and showed that the frequencies of HLA alleles vary not only between Southern African populations and the rest of the world, but that intra-African diversity also exists. The authors show the existence of uniquely African specific alleles from various sources in the literature, and demonstrate the need for additional genetic data from these areas in order to fully understand HLA diversity. Polymorphisms within HLA molecules are thought to have been maintained within populations through selective pressure from exposure to pathogens (Cagliani & Sironi 2013), and in particular, co-evolution between host and pathogen has resulted in *M. tuberculosis* lineages adapting to specific human populations within defined geographical settings (Brites & Gagneux 2015). Certain HLA class I alleles have been correlated with increased susceptibility to TB (Salie *et al*. 2014), with certain HLA alleles associated with disease caused by certain strains from various lineages such as the Euro-American or East Asian lineages. There is therefore the need to investigate both host and pathogen genetic diversity when studying disease development and pathogenesis. The protective effect of a vaccine may be compromised as a result of both host and pathogen genetic variation where clinical trials are conducted on a sample group that does not fully represent the global disease burden. Given the high burden of TB in Southern Africa, and the urgent need for additional vaccines that cater for this population, incorporating information on the distribution of HLA alleles across different populations is crucial for the success of any potential vaccine candidate.

2.4.2 Prediction of potential population coverage

It is evident that when using *in silico* epitope prediction methods, the choice of which HLA alleles to be included in the prediction is an important consideration. Identification of promiscuous epitopes, as well as predicting the potential population coverage is also an important step within a RV workflow.

Various *in silico* epitope prediction tools included in Table 2.4 focus on locating promiscuous binding regions, including ProPred (Singh & Raghava 2001). Given the existence of HLA supertype families, tools such as PEPVAC have been developed to identify peptides able to bind to HLA molecules within a particular supertype family (Reche & Reinherz 2005), and in this way are able to limit the number of potential epitopes without compromising the potential population coverage.

A tool used to predict the potential population coverage of a given set of epitopes is available within the IEDB analysis resource (Kim *et al*. 2012). This algorithm calculates the percentage of individuals expected to respond to a given epitope set based on HLA genotypic frequencies from the dbMHC database (Bui *et al*. 2006). Researchers are also able to create custom populations with associated HLA allele frequencies, which may be useful to estimate the potential protective coverage within a vaccine trial setting. Multiple population coverage's can be aggregated to determine the coverage of a particular set of epitopes across various populations for which a vaccine may be targeted at. This tool can also be used to determine the minimum number of epitope/HLA combinations recognized by 90% of the population.

PREDIVAC is another particularly useful tool which integrates prediction of HLA class II epitopes with population coverage to optimise epitope selection within well-defined ethnic populations (Oyarzún *et al*. 2013). Allele frequencies from the AFND database are used in order to identify vaccine candidates appropriate for a genetically heterogeneous population.

2.5 Identification of Potential TB Vaccine Candidates

Given the vital need for a new vaccine, it is not surprising that RV workflows have been used extensively within *M. tuberculosis* research. Various selection criteria are often applied. These have either focused on a particular family of proteins which are believed to be immunogenic, or broader selection criteria have been imposed, focusing on a specific biological aspect such as secreted proteins or those involved in dormancy or persistence. To address the inherent selection bias of these approaches, whole genome methods also have been applied to the *M. tuberculosis* genome. In either scenario, determining the genetic diversity within epitope-rich regions using data from multiple strains is important and has been used in literature. Evaluation of the protective population coverage has shown promising results for potential candidates as well as those currently in various phases of clinical trials. In addition, various potential candidates predicted via computational tools have been experimentally validated for *M. tuberculosis*.

2.5.1 Selection criteria: family of proteins

An example of a protein family where the function is largely unknown but which are speculated to play a role in host-pathogen interactions is embodied by the PE/PPE family of proteins. In particular, given the high level of sequence variation within the PE_PGRS and PPE_MPTR subfamilies, it has been hypothesized that these proteins play a role in antigenic variation. The PE/PPE proteins and specifically the PE_PGRS proteins have therefore been the focus of various computational analyses involving epitope prediction.

An *in silico* analysis of all PE/PPE proteins was used to predict peptides able to bind to HLA class I molecules and therefore elicit a CD8+ T-cell response (Chaitra *et al*. 2005). An immunoinformatics pipeline was followed which included MHC class I epitope prediction using the BIMAS prediction tool and identification of self-peptides. Molecular modelling and structural analysis was performed using known crystal structures of 5 MHC class I molecules in order to validate binding predictions. A large number of the predicted epitopes had no significant similarities with human peptides, emphasizing the uniqueness of the PE/PPE proteins in mycobacteria, and their potential for being vaccine candidates. Results from this study showed a large number of epitopes in subgroups of the PE/PPE family other than the highly variable PE_PGRS proteins which showed a low level of binding. Peptides bound to alleles from HLA class I B locus had a greater binding affinity than HLA class I A or C locus, and therefore the distribution of HLA B alleles in a population may be an important factor when determining a vaccine's potential protective coverage. Seven peptides predicted to be the strongest binders were experimentally tested. (Chaitra *et al*. 2005; Chaitra, Shaila, *et al*. 2007; Chaitra, Nayak, *et al*. 2007; Chaitra, Shaila, Chandra, *et al*. 2008; Chaitra, Shaila & Nayak 2008). Both CD4+ and CD8+ T-cell epitopes within the PE_PGRS proteins have also been investigated, and particularly whether the genetic diversity within these proteins contributes to human T-cell recognition (Copin *et al*. 2014). Findings from this investigation were consistent with Chaitra *et al*. as the regions containing T-cell epitopes and the genetically diverse regions within the PE_PGRS proteins are distinct, indicating that the most immunogenic PE/PPE proteins are the most conserved of the family. These findings are also consistent with other literature for *M. tuberculosis* which has indicated that T-cell epitopes reported thus far are highly conserved (Comas *et al*. 2010). However, to date, no studies have focused specifically on the PPE_MPTR subfamily which is also highly variable and possibly involved in host-pathogen interactions.

Thus far two PPE proteins, PPE42 and PPE18, have already been included within peptide-based vaccines which have undergone clinical trials. Mtb72F consisting of peptides from pepA as well as PPE18 was the first subunit vaccine for *M. tuberculosis*. Genetic diversity within these genes was investigated using 225 clinical strains from two different geographical locations, and a combination of SNP's and indels resulting in amino acid alterations were found within regions of the PPE18 protein with predicted T-cell epitopes (Hebert *et al*. 2007). The authors concluded that the ability of Mtb72F to induce protective immunity against multiple strains of *M. tuberculosis* may therefore be compromised as a result of the high degree of genetic variation. These findings were recently further investigated using a well-characterized reference collection of 71 *M. tuberculosis* strains from 23 phylogenetic lineages and high variability within the PPE18 gene was confirmed even within the normally conserved N-terminal domain (Homolka *et al*. 2016) These studies yet again highlight the importance of comparative genomics within the RV pipeline for *M. tuberculosis*.

Another example of a protein family that has been the focus of various vaccine investigations using *in silico* techniques is the ESX family, and in particular the ESAT-6 family members, which are often studied in conjunction with the PE/PPE family members given the association of the duplication of the *esx* gene cluster regions with the evolution of the PE/PPE gene family. (Gey van Pittius *et al*. 2006). RV pipelines using biological selection criteria such as growth in macrophages, up or down regulation under hypoxic conditions, secretion or membrane association have also often included the addition of both PE/PPE and *esx* family members (Bertholet *et al*. 2008). A detailed investigation into the vaccine potential of five possible ESX dimer substrates resulted in the identification of a vaccine candidate with high predicted population coverage, and similar protection to BCG when experimentally validated (Knudsen *et al*. 2014). ESAT-6 antigens have also already been included in the *M. tuberculosis* subunit vaccine Ag85B-ESAT-6, which is currently undergoing clinical trials. Given the implications of the genetic variability in PPE18 gene potentially affecting the ability of Mtb72F to induce protective immunity against various clinical strains, genetic variation within the genes encoding ESAT-6 (*esxA* and *esxH*) was also investigated (Davila *et al*. 2010). Within a sample of 88 clinical isolates, no DNA polymorphisms were found in either the *esxA* or *esxH* genes, suggesting that the efficacy of Ag85B-ESAT-6 was unlikely to be affected by genetic diversity within various *M. tuberculosis* populations.

2.5.2 Selection criteria: biological function

Rather than focusing on one particular family of proteins, TB researchers have selected proteins which group together based on a particular biological function. Promiscuous epitopes within proteins associated with dormancy were chosen for HLA class I epitope analysis using ProPred (Sundaramurthi *et al*. 2012). Seven novel promiscuous epitopes, conserved among virulent *M. tuberculosis* strains and showing high predicted population coverage were identified using this approach. In an attempt to study proteins that are highly expressed *in vivo*, Nguyen Thi *et al*. (2014) analysed gene expression profiles of *M. tuberculosis* at various stages of infection, and focused on 38 highly expressed proteins in the active, latent and reactivation phases before predicting both T-cell and B-cell epitopes. A detailed analysis of 742 predicted extracellular or surface localized adhesin and adhesin-like proteins was performed for *M. tuberculosis* and resulted in a set of 233 most probable vaccine candidates which included PE/PPE proteins, Mpt proteins, Cfp2, PstS2, ESAT-6, HBHA, Antigen 85A, Antigen 85B, Antigen 85C as well as many hypothetical proteins (Chaudhuri *et al*. 2014). Analysis of beta-barrel outer-membrane proteins have also identified possible vaccine candidates within the PE/PPE and Mce proteins, as well as other transporters and lipoproteins (Pajón *et al*.).

Secreted proteins have been one of the main focus areas of various *M. tuberculosis* vaccine researchers (Vani *et al*. 2006; De Groot *et al*. 2005). De Groot *et al*. (2005) focused on nine well characterized secreted antigens (antigen 85 complex, Mpt 64, Mpb/Mpt 70, Mpt 63, the 38 kDa protein, the 14 kDa, 16 kDa, 19 kDa, and 32 kDa) and used EpiMer (Meister *et al*. 1995) to identify 23 predicted epitopes within these proteins. In addition, a whole genome approach was used to select an additional 17 predicted epitopes, by selecting open reading frames containing signal sequences using the EpiMatrix tool (Moise *et al*. 2015). Predicted epitopes were experimentally validated by performing gamma-interferon ELISPOT assays using peripheral blood mononuclear cells from latent TB infected individuals. T-cell responses were confirmed for 22 out of the 23 predicted epitopes from EpiMer (96%), and 15 out of 17 (88%) epitopes predicted using EpiMatrix. Epitopes predicted to be promiscuous and shown to induce significant IFN-γ secretion were chosen for inclusion in a prototype vaccine candidate. Of these, one of the most immunogenic epitopes was predicted using the whole genome approach and originated from a protein which at the time of publication, had unknown function and localization. This paper therefore highlights the important role a computational analysis can play in determining vaccine candidates, not only within proteins with known biological relevance but in the absence of functional knowledge of a protein using a whole genome approach.

2.5.3 Whole genome approaches

A whole genome approach allows for an unbiased view of which *M. tuberculosis* proteins may contain potential epitopes and therefore vaccine candidates. Results from a whole genome identification of vaccine candidates by Zvi *et al*. (2008) identified 189 genes with potential epitopes and revealed that the relative distribution of these antigens within 3 different functional categories differed to the distribution of genes within the genome. Prominent representatives within the predicted vaccine candidates again included the PE/PPE and ESX proteins, as well as those involved in dormancy, and proteins localised to the cell wall. Conserved hypothetical proteins also contained potential epitopes. Once again, this highlights that proteins of unknown function may represent an untapped source of vaccine antigens.

2.5.4 Protective coverage:

Assessing the genetic diversity across various *M. tuberculosis* strains should be coupled with assessing the HLA genetic diversity in various human populations in order to determine a vaccine's potential efficacy. This has been investigated for the Mtb72F subunit vaccine (McNamara *et al*. 2010). An *in silico* prediction with different DRB1 genotypes showed that the Mtb72F vaccine would be less effective for several DRB1 genotypes either due to limited epitope binding or to binding to unconserved PPE18 epitopes. Those DRB1 alleles also have a high frequency in high burden TB populations. This study therefore demonstrates that Mtb72F may not be effective for certain host and pathogen genetic circumstances. The predicted population coverage for Ag85B-ASAT-6, Ag85B-TB10.4 and Mtb72f using both HLA class I and II alleles has also been investigated in high TB burden populations (Davila *et al*. 2012). *In silico* methods were used to predict binding of peptide sequences from these vaccine candidates to various alleles. Alleles of concern (alleles predicted to bind four or fewer vaccine epitopes) and populations of concern (populations where >30% of the population were homozygous for an allele of concern) were determined. Results from this study show that the Ag85 vaccine candidates have superior potential population coverage than the Mtb72f vaccine candidate. Four HLA class I alleles did not bind to any of the Mtb72f epitopes; with these alleles among the three most prevalent alleles in 7 out of the 22 high burden TB countries defined by WHO.

2.5.5 Experimental validation:

Potential vaccine candidates identified from RV pipelines ultimately need to be validated. Various studies discussed in this review have used wet lab testing to validate the results from *in silico* predictions. Panigada (2002) selected mycobacterial cell entry genes to investigate for possible CD4+ T-cell epitopes using the TEPITOPE tool. Five peptides predicted to be potential HLA-DR ligands were tested for induction of proliferation of CD4+ cells experimentally. This led to the identification of a peptide able to induce CD4+ cell proliferation in 50% of the tested subjects. After using a transcriptional based selection criteria, Gideon *et al*. (2012) evaluated the prediction results from ProPred using IFN-γ and IL-2 ELISPOT assays. Twenty three of twenty six proteins induced an IFN-γ response and five novel immunodominant proteins were identified. A whole genome based approach was used to identify 432 potential CD8+ T-cell epitopes able to bind to 3 HLA alleles each representing one of the HLA supertype families, and peptide/HLA binding affinity measured *in vitro*, as well as CD8+ T-cell proliferation assays and intracellular cytokine staining for IFN-γ, IL-2 and TNF-α used to validate results (Tang *et al*. 2011). Seventy out of the 432 epitopes were confirmed, 58 of which were novel and have not previously been described by other studies.

2.6 Conclusion

There is a crucial need for a new vaccine against *M. tuberculosis,* and subunit vaccines have shown promising results thus far. *In silico* methods are allowing TB researchers to search the *M. tuberculosis* proteome for epitopes at a much faster rate than ever before. Experimental designs can now be guided and informed by the results from an *in silico* analysis, where the starting point is the most likely outcome, instead of being guided by an incomplete current biological understanding. Immunoinformatics, and reverse vaccinology are continuously improving and the tools becoming more accurate and reliable, especially given the growing amount of data being generated daily used to train the predictive algorithms.

Given the high HLA diversity in the human population as well as genetic variation across strains, comparative genomics and prediction of protective coverage is extremely important for a pathogen such as *M. tuberculosis*. Immunoinformatics therefore has the potential to play a large role in vaccine trial planning and expected success of the project. Host-customised vaccines where the genetic background of the host is taken into account before the most appropriate vaccine cocktail is administered may also be a viable option in the future. With time and costs associated with genome sequencing decreasing, not only is host directed drug prescription as in the case of pharmacogenetics becoming a reality, but the same could be applied for the administration of vaccines. This may overcome the problem of protective immunity being compromised by variation in HLA molecules in specific populations.

Many studies from various groups have focused on the PE/PPE family of proteins in their search for potential epitopes. In addition, resulting epitopes from whole genome approaches consistently contain at least a subset of the PE/PPE proteins within them. The PPE_MPTR protein family however has thus far been left out of most investigations, due to their largely unknown function, high genetic variation and difficulty in working with them in an experimental setting. The PPE_MPTR proteins are therefore an ideal target for an immunoinformatic based approach.

2.7 References

Bambini, S. & Rappuoli, R., 2009. The use of genomics in microbial vaccine development. *Drug Discovery Today*, 14(5-6), pp.252–260.

Bertholet, S. *et al*., 2008. Identification of human T cell antigens for the development of vaccines against *Mycobacterium tuberculosis*. *Journal of immunology (Baltimore, Md. : 1950)*, 181(11), pp.7948–57.

Bhasin, M. & Raghava, G.P.S., 2007. A hybrid approach for predicting promiscuous MHC class I restricted T cell epitopes. *Journal of biosciences*, 32(1), pp.31–42.

Bhasin, M. & Raghava, G.P.S., 2004a. Prediction of CTL epitopes using QM, SVM and ANN techniques. *Vaccine*, 22(23-24), pp.3195–204.

Bhasin, M. & Raghava, G.P.S., 2004b. SVM based method for predicting HLA-DRB1*0401 binding peptides in an antigen sequence. *Bioinformatics (Oxford, England)*, 20(3), pp.421–3.

Bian, H. & Hammer, J., 2004. Discovery of promiscuous HLA-II-restricted T cell epitopes with TEPITOPE. *Methods (San Diego, Calif.)*, 34(4), pp.468–75.

Blythe, M.J. *et al*., 2007. An analysis of the epitope knowledge related to Mycobacteria. *Immunome Research*, 3(1), p.10.

Blythe, M.J., Doytchinova, I.A. & Flower, D.R., 2002. JenPep: a database of quantitative functional peptide data for immunology. *Bioinformatics (Oxford, England)*, 18(3), pp.434–9.

Breuer, K. *et al*., 2013. InnateDB: systems biology of innate immunity and beyond--recent updates and continuing curation. *Nucleic acids research*, 41(Database issue), pp.D1228–33.

Brites, D. & Gagneux, S., 2015. Co-evolution of *Mycobacterium tuberculosis* and Homo sapiens. *Immunological reviews*, 264(1), pp.6–24.

Brusic, V. & Petrovsky, N., 2005. Immunoinformatics and its relevance to understanding human immune disease. *Expert review of clinical immunology*, 1(1), pp.145–57.

Bui, H.-H. *et al*., 2005. Automated generation and evaluation of specific MHC binding predictive tools: ARB matrix applications. *Immunogenetics*, 57(5), pp.304–14.

Bui, H.-H. *et al*., 2006. Predicting population coverage of T-cell epitope-based diagnostics and vaccines. *BMC bioinformatics*, 7, p.153.

Cagliani, R. & Sironi, M., 2013. Pathogen-driven selection in the human genome. *International journal of evolutionary biology*, 2013, p.204240.

Chaitra, M.G. *et al*., 2005. Defining putative T cell epitopes from PE and PPE families of proteins of *Mycobacterium tuberculosis* with vaccine potential. *Vaccine*, 23(10), pp.1265–1272.

Chaitra, M.G., Shaila, M.S., Chandra, N.R., *et al*., 2008. HLA-A*0201-restricted cytotoxic T-cell epitopes in three PE/PPE family proteins of *Mycobacterium tuberculosis*. *Scandinavian Journal of Immunology*, 67(4), pp.411–417.

Chaitra, M.G., Nayak, R. & Shaila, M.S., 2007. Modulation of immune responses in mice to recombinant antigens from PE and PPE families of proteins of *Mycobacterium tuberculosis* by the Ribi adjuvant. *Vaccine*, 25(41), pp.7168–7176.

Chaitra, M.G., Shaila, M.S. & Nayak, R., 2008. Detection of interferon gamma-secreting CD8+ T lymphocytes in humans specific for three PE/PPE proteins of *Mycobacterium tuberculosis*. *Microbes and Infection*, 10(8), pp.858–867.

Chaitra, M.G., Shaila, M.S. & Nayak, R., 2007. Evaluation of T-cell responses to peptides with MHC class I-binding motifs derived from PE_PGRS 33 protein of *Mycobacterium tuberculosis*. *Journal of Medical Microbiology*, 56(4), pp.466–474.

Chaudhuri, R. *et al*., 2014. Integrative immunoinformatics for Mycobacterial diseases in R platform. *Systems and synthetic biology*, 8(1), pp.27–39.

Cole, S.T. *et al*., 1998. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature*, 393(6685), pp.537–44.

Comas, I. *et al*., 2010. Human T cell epitopes of *Mycobacterium tuberculosis* are evolutionarily hyperconserved. *Nature genetics*, 42(6), pp.498–503.

Copin, R. *et al*., 2014. Sequence diversity in the pe_pgrs genes of *Mycobacterium tuberculosis* is independent of human T cell recognition. *mBio*, 5(1), pp.e00960–13.

Cui, J. *et al*., 2006. MHC-BPS: MHC-binder prediction server for identifying peptides of flexible lengths from sequence-derived physicochemical properties. *Immunogenetics*, 58(8), pp.607–13.

Daniel, S. *et al*., 1998. Relationship between peptide selectivities of human transporters associated with antigen processing and HLA class I molecules. *Journal of immunology (Baltimore, Md. : 1950)*, 161(2), pp.617–24.

Davies, M.N. & Flower, D.R., 2007. Harnessing bioinformatics to discover new vaccines. *Drug Discovery Today*, 12(9-10), pp.389–395.

Davila, J. *et al*., 2010. Assessment of the genetic diversity of *Mycobacterium tuberculosis* esxA, esxH, and fbpB genes among clinical isolates and its implication for the future immunization by new tuberculosis subunit vaccines Ag85B-ESAT-6 and Ag85B-TB10.4. *Journal of biomedicine & biotechnology*, 2010, p.208371.

Davila, J., McNamara, L.A. & Yang, Z., 2012. Comparison of the predicted population coverage of tuberculosis vaccine candidates Ag85B-ESAT-6, Ag85B-TB10.4, and Mtb72f via a bioinformatics approach. *PloS one*, 7(7), p.e40882.

DeLuca, D.S., Khattab, B. & Blasczyk, R., 2007. A modular concept of HLA for comprehensive peptide binding prediction. *Immunogenetics*, 59(1), pp.25–35.

Dimitrov, I. *et al*., 2010. EpiTOP--a proteochemometric tool for MHC class II binding prediction. *Bioinformatics (Oxford, England)*, 26(16), pp.2066–8.

Dönnes, P. & Kohlbacher, O., 2005. Integrated modeling of the major events in the MHC class I antigen processing pathway. *Protein science : a publication of the Protein Society*, 14(8), pp.2132–40.

Dönnes, P. & Kohlbacher, O., 2006. SVMHC: a server for prediction of MHC-binding peptides. *Nucleic acids research*, 34(Web Server issue), pp.W194–7.

Doytchinova, I.A., Guan, P. & Flower, D.R., 2006. EpiJen: a server for multistep T cell epitope prediction. *BMC bioinformatics*, 7, p.131.

Galagan, J.E. *et al*., 2010. TB database 2010: overview and update. *Tuberculosis (Edinburgh, Scotland)*, 90(4), pp.225–35.

Gey van Pittius, N.C. *et al*., 2006. *Evolution and expansion of the Mycobacterium tuberculosis PE and PPE multigene families and their association with the duplication of the ESAT-6 (esx) gene cluster regions.*,

Gideon, H.P. *et al*., 2012. Bioinformatic and Empirical Analysis of Novel Hypoxia-Inducible Targets of the Human Antituberculosis T Cell Response. *The Journal of Immunology*, 189(12), pp.5867–5876.

González-Galarza, F.F. *et al*., 2015. Allele frequency net 2015 update: new features for HLA epitopes, KIR and disease and HLA adverse drug reaction associations. *Nucleic acids research*, 43(Database issue), pp.D784–8.

Greenbaum, J. *et al*., 2011. Functional classification of class II human leukocyte antigen (HLA) molecules reveals seven different supertypes and a surprising degree of repertoire sharing across supertypes. *Immunogenetics*, 63(6), pp.325–35.

Grode, L. *et al*., 2005. Increased vaccine efficacy against tuberculosis of recombinant Mycobacterium bovis bacille Calmette-Guérin mutants that secrete listeriolysin. *The Journal of clinical investigation*, 115(9), pp.2472–9.

De Groot, A.S. *et al*., 2005. Developing an epitope-driven tuberculosis (TB) vaccine. *Vaccine*, 23(17-18), pp.2121–2131.

Guan, P. *et al*., 2006. MHCPred 2.0: an updated quantitative T-cell epitope prediction server. *Applied bioinformatics*, 5(1), pp.55–61.

Hakenberg, J. *et al*., 2003. MAPPP: MHC class I antigenic peptide processing prediction. *Applied bioinformatics*, 2(3), pp.155–8.

Hebert, A.M. *et al*., 2007. DNA polymorphisms in the pepA and PPE18 genes among clinical strains of *Mycobacterium tuberculosis*: implications for vaccine efficacy. *Infection and immunity*, 75(12), pp.5798–805.

Helmberg, W., Dunivin, R. & Feolo, M., 2004. The sequencing-based typing tool of dbMHC: typing highly polymorphic gene sequences. *Nucleic acids research*, 32(Web Server issue), pp.W173–5.

Hertz, T. & Yanover, C., 2006. PepDist: a new framework for protein-peptide binding prediction based on learning peptide distance functions. *BMC bioinformatics*, 7 Suppl 1, p.S3.

Homolka, S., Ubben, T. & Niemann, S., 2016. High Sequence Variability of the ppE18 Gene of Clinical *Mycobacterium tuberculosis* Complex Strains Potentially Impacts Effectivity of Vaccine Candidate M72/AS01E. *PloS one*, 11(3), p.e0152200.

Jojic, N. *et al*., 2006. Learning MHC I--peptide binding. *Bioinformatics*, 22(14), pp.e227–e235.

Karosiene, E. *et al*., 2013. NetMHCIIpan-3.0, a common pan-specific MHC class II prediction method including all three human MHC class II isotypes, HLA-DR, HLA-DP and HLA-DQ. *Immunogenetics*, 65(10), pp.711–24.

Kim, Y. *et al*., 2012. Immune epitope database analysis resource. *Nucleic acids research*, 40(Web Server issue), pp.W525–30.

Knudsen, N.P.H. *et al*., 2014. Tuberculosis vaccine with high predicted population coverage and compatibility with modern diagnostics. *Proceedings of the National Academy of Sciences of the United States of America*, 111(3), pp.1096–101.

Kobayashi, K.S. & van den Elsen, P.J., 2012. NLRC5: a key regulator of MHC class I-dependent immune responses. *Nature reviews. Immunology*, 12(12), pp.813–20.

Larsen, M.V. *et al*., 2005. An integrative approach to CTL epitope prediction: a combined algorithm integrating MHC class I binding, TAP transport efficiency, and proteasomal cleavage predictions. *European journal of immunology*, 35(8), pp.2295–303.

Lata, S., Bhasin, M. & Raghava, G.P.S., 2009. MHCBN 4.0: A database of MHC/TAP binding peptides and T-cell epitopes. *BMC research notes*, 2, p.61.

Lefranc, M.-P., 2008. IMGT, the International ImMunoGeneTics Information System for Immunoinformatics : methods for querying IMGT databases, tools, and web resources in the context of immunoinformatics. *Molecular biotechnology*, 40(1), pp.101–11.

Lew, J.M. *et al*., 2011. TubercuList--10 years after. *Tuberculosis (Edinburgh, Scotland)*, 91(1), pp.1–7.

Lin, H.H., Ray, S., *et al*., 2008. Evaluation of MHC class I peptide binding prediction servers: applications for vaccine research. *BMC immunology*, 9, p.8.

Lin, H.H., Zhang, G.L., *et al*., 2008. Evaluation of MHC-II peptide binding prediction servers: applications for vaccine research. *BMC bioinformatics*, 9 Suppl 12, p.S22.

Longmate, J. *et al*., 2001. Population coverage by HLA class-I restricted cytotoxic T-lymphocyte epitopes. *Immunogenetics*, 52(3-4), pp.165–73.

Loxton, A., Hondalus, M. & Sampson, S., 2016. TB vaccine assessment. In P. B. F. Anthony J. Hickey, Amit Misra, ed. *Delivery Systems for Tuberculosis Prevention and Treatment*. Wiley, pp. pp91–110.

Lundegaard, C. *et al*., 2008. NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8-11. *Nucleic Acids Research*, 36(Web Server), pp.W509–W512.

McNamara, L. a, He, Y. & Yang, Z., 2010. Using epitope predictions to evaluate efficacy and population coverage of the Mtb72f vaccine for tuberculosis. *BMC immunology*, 11, p.18.

McSparron, H. *et al*., 2003. JenPep: a novel computational information resource for immunobiology and vaccinology. *Journal of chemical information and computer sciences*, 43(4), pp.1276–87.

Meister, G.E. *et al*., 1995. Two novel T cell epitope prediction algorithms based on MHC-binding motifs; comparison of predicted and published epitopes from *Mycobacterium tuberculosis* and HIV protein sequences. *Vaccine*, 13(6), pp.581–91.

Moise, L. *et al*., 2015. iVAX: An integrated toolkit for the selection and optimization of antigens and the design of epitope-driven vaccines. *Human vaccines & immunotherapeutics*, 11(9), pp.2312–21.

Monterrubio-López, G.P., González-Y-Merchand, J.A. & Ribas-Aparicio, R.M., 2015. Identification of Novel Potential Vaccine Candidates against Tuberculosis Based on Reverse Vaccinology. *BioMed research international*, 2015, p.483150.

Mora, M. *et al*., 2006. Microbial genomes and vaccine design: refinements to the classical reverse vaccinology approach. *Current Opinion in Microbiology*, 9(5), pp.532–536.

Moutaftsi, M. *et al*., 2006. A consensus epitope prediction approach identifies the breadth of murine T(CD8+)-cell responses to vaccinia virus. *Nature biotechnology*, 24(7), pp.817–9.

Movahedi, A.R. & Hampson, D.J., 2008. New ways to identify novel bacterial antigens for vaccine development. *Veterinary microbiology*, 131(1-2), pp.1–13.

Nguyen Thi, L.T. *et al*., 2014. Immunoinformatics study on highly expressed *Mycobacterium tuberculosis* genes during infection. *Tuberculosis*, 94(5), pp.475–481.

Nielsen, M. *et al*., 2010. MHC class II epitope predictive algorithms. *Immunology*, 130(3), pp.319–28.

Nielsen, M. *et al*., 2003. Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. *Protein science : a publication of the Protein Society*, 12(5), pp.1007–17.

Nielsen, M. & Lund, O., 2009. NN-align. An artificial neural network-based alignment algorithm for MHC class II peptide binding prediction. *BMC bioinformatics*, 10, p.296.

Nielsen, M., Lundegaard, C. & Lund, O., 2007. Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method. *BMC bioinformatics*, 8, p.238.

Oyarzún, P. *et al*., 2013. PREDIVAC: CD4+ T-cell epitope prediction for vaccine design that covers 95% of HLA class II DR protein diversity. *BMC bioinformatics*, 14, p.52.

Pajón, R. *et al*., Computational identification of beta-barrel outer-membrane proteins in *Mycobacterium tuberculosis* predicted proteomes as putative vaccine candidates. *Tuberculosis (Edinburgh, Scotland)*, 86(3-4), pp.290–302.

Panigada, M., 2002. Identification of a Promiscuous T-Cell Epitope in *Mycobacterium tuberculosis* Mce Proteins. *Infection and Immunity*, 70(1), pp.79–85.

Pappalardo, F. *et al*., 2009. ImmunoGrid, an integrative environment for large-scale simulation of the immune system for vaccine discovery, design and optimization. *Briefings in bioinformatics*, 10(3), pp.330–40.

Parker, K.C., Bednarek, M.A. & Coligan, J.E., 1994. Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains. *Journal of immunology (Baltimore, Md. : 1950)*, 152(1), pp.163–75.

Paximadis, M. *et al*., 2012. Human leukocyte antigen class I (A, B, C) and II (DRB1) diversity in the black and Caucasian South African population. *Human immunology*, 73(1), pp.80–92.

Peters, B. *et al*., 2003. Examining the independent binding assumption for binding of peptide epitopes to MHC-I molecules. *Bioinformatics (Oxford, England)*, 19(14), pp.1765–72.

Peters, B. & Sette, A., 2005. Generating quantitative models describing the sequence specificity of biological processes with the stabilized matrix method. *BMC bioinformatics*, 6, p.132.

Pizza, M. *et al*., 2000. Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing. *Science (New York, N.Y.)*, 287(5459), pp.1816–1820.

Rammensee, H. *et al*., 1999. SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics*, 50(3-4), pp.213–9.

Rapin, N. *et al*., 2010. Computational immunology meets bioinformatics: the use of prediction tools for molecular binding in the simulation of the immune system. *PloS one*, 5(4), p.e9862.

Reche, P.A. *et al*., 2005. EPIMHC: a curated database of MHC-binding peptides for customized computational vaccinology. *Bioinformatics (Oxford, England)*, 21(9), pp.2140–1.

Reche, P.A., Glutting, J.-P. & Reinherz, E.L., 2002. Prediction of MHC class I binding peptides using profile motifs. *Human immunology*, 63(9), pp.701–9.

Reche, P.A. & Reinherz, E.L., 2005. PEPVAC: a web server for multi-epitope vaccine development based on the prediction of supertypic MHC ligands. *Nucleic acids research*, 33(Web Server issue), pp.W138–42.

Rezwan, M. *et al*., 2007. Lipoprotein synthesis in mycobacteria. *Microbiology (Reading, England)*, 153(Pt 3), pp.652–8.

Robinson, J. *et al*., 2014. The IPD and IMGT/HLA database: allele variant databases. *Nucleic acids research*, p.gku1161–.

Salie, M. *et al*., 2014. Associations between human leukocyte antigen class i variants and the *Mycobacterium tuberculosis* subtypes causing disease. *Journal of Infectious Diseases*, 209(2), pp.216–223.

Serruto, D. *et al*., 2009. Genome-based approaches to develop vaccines against bacterial pathogens. *Vaccine*,

27(25-26), pp.3245–3250.

Sidney, J. *et al*., 2008. Quantitative peptide binding motifs for 19 human and mouse MHC class I molecules derived using positional scanning combinatorial peptide libraries. *Immunome research*, 4, p.2.

Singh, H. & Raghava, G.P., 2001. ProPred: prediction of HLA-DR binding sites. *Bioinformatics (Oxford, England)*, 17(12), pp.1236–7.

Singh, H. & Raghava, G.P.S., 2003. ProPred1: prediction of promiscuous MHC Class-I binding sites. *Bioinformatics (Oxford, England)*, 19(8), pp.1009–14.

Sundaramurthi, J.C. *et al*., 2012. *In silico* identification of potential antigenic proteins and promiscuous CTL epitopes in *Mycobacterium tuberculosis*. *Infection, Genetics and Evolution*, 12(6), pp.1312–1318.

Tang, S.T. *et al*., 2011. Genome-based *in silico* identification of new *Mycobacterium tuberculosis* antigens activating polyfunctional CD8+ T cells in human tuberculosis. *Journal of immunology (Baltimore, Md. : 1950)*, 186(2), pp.1068–1080.

Taylor, P.D. *et al*., 2006. LIPPRED: A web server for accurate prediction of lipoprotein signal sequences and cleavage sites. *Bioinformation*, 1(5), pp.176–9.

Tomar, N. & De, R.K., 2014. Immunoinformatics: a brief review. *Methods in molecular biology (Clifton, N.J.)*, 1184, pp.23–55.

Tomar, N. & De, R.K., 2010. Immunoinformatics: an integrated scenario. *Immunology*, 131(2), pp.153–68.

Toseland, C.P. *et al*., 2005. AntiJen: a quantitative immunology database integrating functional, thermodynamic, kinetic, biophysical, and cellular data. *Immunome research*, 1(1), p.4.

Tshabalala, M., Mellet, J. & Pepper, M.S., 2015. Human Leukocyte Antigen Diversity: A Southern African Perspective. , 2015(class I).

Vani, J. *et al*., 2006. A combined immuno-informatics and structure-based modeling approach for prediction of T cell epitopes of secretory proteins of *Mycobacterium tuberculosis*. *Microbes and Infection*, 8(3), pp.738–746.

Vita, R. *et al*., 2014. The immune epitope database (IEDB) 3.0. *Nucleic Acids Research*, 43(D1), pp.D405–D412.

Vivona, S. *et al*., 2008. Computer-aided biotechnology: from immuno-informatics to reverse vaccinology. *Trends in Biotechnology*, 26(4), pp.190–200.

Vivona, S., Bernante, F. & Filippini, F., 2006. NERVE: new enhanced reverse vaccinology environment. *BMC biotechnology*, 6(1), p.35.

Wang, P. *et al*., 2008. A systematic assessment of MHC class II peptide binding predictions and evaluation of a consensus approach. *PLoS computational biology*, 4(4), p.e1000048.

Wang, P. *et al*., 2010. Peptide binding predictions for HLA DR, DP and DQ molecules. *BMC bioinformatics*, 11(1), p.568.

WHO, 2015. *Global tuberculosis report 2015*, World Health Organization.

Zhang, G.L. *et al*., 2011. Dana-Farber repository for machine learning in immunology. *Journal of immunological methods*, 374(1-2), pp.18–25.

Zhang, G.L. *et al*., 2005. MULTIPRED: a computational system for prediction of promiscuous HLA binding peptides. *Nucleic acids research*, 33(Web Server issue), pp.W172–9.

Zhang, G.L. *et al*., 2006. PRED(TAP): a system for prediction of peptide binding to the human transporter associated with antigen processing. *Immunome research*, 2, p.3.

Zhang, L. *et al*., 2012. TEPITOPEpan: extending TEPITOPE for peptide binding prediction covering over 700 HLA-DR molecules. *PloS one*, 7(2), p.e30483.

Zvi, A. *et al*., 2008. Whole genome identification of *Mycobacterium tuberculosis* vaccine candidates by comprehensive data mining and bioinformatic analyses. *BMC medical genomics*, 1, p.18.

Chapter 3: Evaluation of MHC II Epitope Prediction Tools

3.1 Introduction

An essential step in CD4+ T-cell activation is the recognition of epitopes presented by HLA class II molecules on the surface of antigen presenting cells (such as macrophages and dendritic cells) by T-cell receptors. In order for this to occur, an antigen peptide (derived from the degradation of internalized pathogen proteins) must first bind to the binding groove of an HLA class II molecule before being transported to the cell surface (Chapter 2, Figure 2.1). HLA molecules are highly polymorphic, with the majority of the polymorphisms reported within this binding groove. Binding of a peptide to an HLA class II molecule is therefore highly specific, where an antigen peptide may only bind to one, or in the case of promiscuous epitopes, a few specific HLA alleles. The majority of epitope prediction tools indirectly predict epitopes by predicting the binding affinity of antigen peptides to various HLA molecules.

There are currently more than 20 different MHC class II prediction tools publically available. The accuracy of these tools varies based on the prediction method used (Chapter 2, section 2.3.1), as well as on the amount of experimental data which has been used to train the prediction algorithm.

In order to determine the most appropriate tool to be used for the prediction of *Mycobacterium tuberculosis* epitopes within the PPE_MPTR proteins, we have performed an evaluation of 8 MHC II epitope prediction tools using known *M. tuberculosis* epitopes. Previous evaluations of these tools can be found in literature (Lin *et al*. 2008; Wang *et al*. 2008), which are based on epitope data from a collection of many source antigens. The unique characteristics of the *M. tuberculosis* genome, including the higher GC content compared to other bacteria, IS*6110* insertion elements, and repeat regions, may influence peptide composition and therefore epitopes from *M. tuberculosis* may display dissimilar binding patterns to HLA alleles compared to epitopes from other pathogens. In particular, the peptide compositions of the gly-ala-asn rich PPE_MPTR proteins may be considerably different from other source peptides used in previous evaluations. Prediction methods are trained using known epitope data from a variety of pathogens, and the reported accuracy of the tools is heavily dependent on the amount and type of training data used. More specifically, a high degree of sequence similarity in the test and training data sets, may lead to overly optimistic estimates of the performance of the tools (El-Manzalawy *et al*. 2008). Therefore if it is indeed the case that the binding specificity of *M. tuberculosis* epitopes differs considerably from other pathogens and an insufficient amount of representative *M. tuberculosis* epitope data is included in the training data set, the accuracy of predictions using these tools may differ for *M. tuberculosis* antigens compared to other source antigens. It is therefore important to distinguish whether the results of such an evaluation would differ when using only *M. tuberculosis* proteins rather than a collection of proteins from various source pathogens.

3.2 Methods

3.2.1 Epitope data

A collection of positive and negative *M. tuberculosis* epitopes was retrieved from the Immune Epitope Database (IEDB), accessed 27/05/2015. Search and filtering criteria are shown in Table 3.1.

Table 3.1: IEDB search criteria. Filtering criteria used to search for *M. tuberculosis* epitopes are shown.

Data Column	Criteria
Source Organism	*M. tuberculosis*
Host	Human
Assay Type	MHC Ligand Assays
MHC allele class	II
Method/Technique	purified MHC/competitive/radioactivity
Assay Group	dissociation constant KD (~IC50)
Result	Positive and Negative

Resulting data included MHC ligand assay results in the form of binding affinities (IC50 nM) from 123 unique peptides (15 mers) from 9 *M. tuberculosis* antigens (Table 3.2) for 30 different HLA class II alleles (6 DPA1/DPB1, 6 DQA1/DQB1, 18 DRB) (Table 3.3). Not every unique peptide had a binding result for all 30 HLA alleles included in the data. For certain peptides, binding results for only a limited number of HLA alleles were available. In total, the data set contained results for 3,343 peptide/HLA combinations.

Table 3.2: *M. tuberculosis* antigens included in IEDB evaluation data. For each antigen, the number of unique 15 mer peptides and the total number of HLA class II ligand assay results is shown. Each unique peptide was tested for binding for up to 30 different HLA molecules.

M. tuberculosis Antigen	Number of Unique Peptides	Total number of assay results
14 kDa antigen	29	787
6 kDa early secretory antigenic target	22	604
Antitoxin Rv2654c/MT2731	15	405
ESAT-6-like protein EsxB	21	573
ESAT-6-like protein EsxJ	19	515
ESAT-6-like protein EsxK	3	81
PPE FAMILY PROTEIN	12	324
Putative ESAT-6-like protein 10	1	27
Putative ESAT-6-like protein 7	1	27
Total	**123**	**3343**

Table 3.3: HLA alleles included in evaluation data. For each HLA allele, the number of *M. tuberculosis* peptides (15 mers) that were tested for binding is shown.

HLA Alleles	Number of Peptides tested against
DPA1/DPB1	**615**
HLA-DPA1*01:03/DPB1*02:01	112
HLA-DPA1*01:03/DPB1*03:01	11
HLA-DPA1*01:03/DPB1*04:01	123
HLA-DPA1*02:01/DPB1*01:01	123
HLA-DPA1*02:01/DPB1*05:01	123
HLA-DPA1*03:01/DPB1*04:02	123
DQA1/DQB1	**738**
HLA-DQA1*01:01/DQB1*05:01	123
HLA-DQA1*01:02/DQB1*06:02	123
HLA-DQA1*03:01/DQB1*03:02	123
HLA-DQA1*04:01/DQB1*04:02	123
HLA-DQA1*05:01/DQB1*02:01	123
HLA-DQA1*05:01/DQB1*03:01	123
DRB	**1990**
HLA-DRB1*01:01	123
HLA-DRB1*03:01	123
HLA-DRB1*04:01	123
HLA-DRB1*04:04	123
HLA-DRB1*04:05	123
HLA-DRB1*07:01	123
HLA-DRB1*08:02	123
HLA-DRB1*09:01	123
HLA-DRB1*10:01	11
HLA-DRB1*11:01	123
HLA-DRB1*12:01	123
HLA-DRB1*13:02	123
HLA-DRB1*15:01	123
HLA-DRB1*16:02	11
HLA-DRB3*01:01	123
HLA-DRB3*02:02	123
HLA-DRB4*01:01	123
HLA-DRB5*01:01	123
Total Peptide:HLA Allele Combinations	**3343**

3.2.2 Tools evaluated

Eight prediction tools (Table 3.4) were used to predict the binding affinity of each of the peptide/HLA combinations contained in the data described above. The tools chosen included those endorsed by the IEDB analysis resource (Kim *et al*. 2012), as well as those with the best results from previous evaluations found in literature (Lin *et al*. 2008; Wang *et al*. 2008). ProPred and TEPITOPE tools both make use of the Sturniolo database and prediction algorithm, and therefore results for these tools were identical. The IEDB consensus tool uses the prediction results from four tools (NN-align, SMM-align, Comblib and Sturniolo) and outputs a median value. Certain tools contain only a limited number of HLA alleles for which prediction is possible and therefore it was not possible to test all peptide/HLA allele combinations for certain tools.

Table 3.4: Prediction tools included in the valuation. The prediction method used within each tool is shown. For each tool, the number of HLA alleles for which prediction is possible out of the alleles included in the IEDB data is shown.

Tool	Method	HLA alleles			Reference
		DPA1/DPB1	DQA1/DQB1	DRB	
NetMHCII	ANN	5/6	6/6	14/18	(Lundegaard *et al*. 2008)
NetMHCIIpan	ANN & SMM	6/6	6/6	18/18	(Karosiene *et al*. 2013)
ARB	ANN & SMM	4/6	5/6	14/18	(Bui *et al*. 2005)
IEDB Consensus	Consensus	4/6	6/6	15/18	(Wang *et al*. 2008)
NN-align	ANN	5/6	6/6	14/18	(Nielsen & Lund 2009)
SMM-align	SMM	4/6	6/6	15/18	(Nielsen *et al*. 2007)
Comblib	Matrix	4/6	6/6	5/18	(Wang *et al*. 2010)
Sturniolo (Propred/Tepitope)	Matrix	0/6	0/6	11/18	(Sturniolo *et al*. 1999; Zhang *et al*. 2012; Singh & Raghava 2001)

3.2.3 Interpretation of prediction results

Prediction results from different tools are given in various formats, including predicted IC50 values, raw relative scores or median percentile ranks. In order to compare results between the various tools, it is necessary to be able to determine comparable cut-off criteria to distinguish between binders and non-binders.

The predicted results for NetMHCII, NetMHCIIpan, ARB, NN-align, SMM-align and Comblib are given in units of IC50 nM, with a lower value indicating a higher binding affinity and therefore a more likely T-cell epitope. These values can be directly compared to the actual IC50 nM values in the IEDB data for each peptide/HLA allele combination. IEDB guidelines recommend using IC50 nM < 50 as high affinity binding; 50 ≤ IC50 nM < 500 as intermediate affinity binding, 500 ≤ IC50 nM < 5000 as low affinity binding; and IC50 NM > 5000 as a non-binder and therefore not a predicted epitope (Kim *et al*. 2012). For the purpose of the evaluation study, a strict and intermediate cut-off value of <100 and <1000 respectively was used to determine positive binders and therefore predicted epitopes. Given the limited size of the data set used in this evaluation, choosing a very low IC50 value as a cut-off would substantially decrease the number of positive binders making any statistical interpretation of the accuracy of the prediction results unreliable. Therefore a strict cut-off value of 100 nM rather than 50 nM was used for the purpose of the valuation. However when using the tools to determine epitopes, a lower cut-off value should be used.

The predicted results for Sturniolo (ProPred/TEPITOPE) are given as a raw score, with higher scores indicating higher affinity binding. Scores from this analysis ranged between -12.5 to 9. A strict and intermediate cut-off value of >1 and >0 respectively was used to determine positive binders and therefore predicted epitopes.

For the IEDB consensus tool, prediction results are in the form of a percentile rank which is generated for each of the four methods used within the consensus method (NN-align, SMM-align, Comblib and Sturniolo), and a median percentile rank is given. The percentile rank is found by comparing a peptide's binding score against the scores of five million random 15 mers selected from the SWISSPROT database. A small percentile rank indicates high binding affinity. The IEDB recommends using a score of <10 for binders and therefore predicted epitopes. A strict and intermediate cut-off value of <10 and <50 respectively was used to determine positive binders and therefore predicted epitopes.

The cut-off values for Sturniolo and IEDB consensus were chosen by comparing the distribution of predicted results to that of the predicted IC50 nM values. Based on this, a value of 100 nM was comparable to a Sturniolo score of 1 and an IEDB consensus percentile rank of 10, and a value of 1000 nM was comparable to a Sturniolo score of 0 and an IEDB consensus percentile rank of 50.

The above mentioned cut-off values were used to count the number of predicted binders and non-binders from each tool. These were compared to the actual number of binders and non-binders in the data. Accuracy, sensitivity and specificity were calculated for each tool.

3.2.4 Accuracy by HLA allele

Previous studies have shown that the accuracy of specific tools varies by HLA allele (Lin *et al*. 2008). Accuracy statistics for each tool were therefore compared individually for each HLA allele. This was only done for HLA alleles for which at least 20 positive data points was available in the data, when using the medium cut-off values. For each allele, the tool with the highest accuracy, sensitivity, and specificity using the medium cut-off value was determined. The IEDB have undergone a similar investigation, where the best tool for each allele has been determined (Kim *et al*. 2012). Results from this evaluation were compared to the IEDB recommended tool for each HLA allele.

The accuracy of the tools is expected to differ since different tools are based on different prediction algorithms and the amount of data that has been used to train the prediction methods may vary. If a specific HLA allele is not found within the training data, predicted binding to that HLA allele will not be available for that tool. However pan methods such as NetMHCIIpan have been created to make predictions for those HLA alleles for which no training data is available. These predictions are based on the similarity of the binding groove to those alleles for which training data is available. The accuracy of the results for these alleles is therefore expected to be lower than for alleles for which actual training data is available. Results for NetMHCIIpan is therefore split between alleles that are included in NetMHCII and training data is available (known) and others which are based on similarity of the binding grooves (unknown).

3.3 Results

3.3.1 Overall accuracy, sensitivity and specificity

The overall accuracy, sensitivity and specificity (averaged across all HLA alleles included in the evaluation study) are shown in Figure 3.1 for both the strict and medium cut-off values.

Strict cut-off values resulted in a lower number of positive data points. Therefore a lower sensitivity and higher specificity for the strict cut-off may be indicative of the smaller amount of positive data rather than a higher number of false predictions at stricter cut-offs. Based on the medium cut-off values, the tools with the highest accuracy are NN-align and NetMHCII with 79.99%.The IEDB consensus tool has the highest sensitivity at 92.49%, but the lowest specificity at 43.19%. Similarly, the SMM-align tool has the highest specificity at 88.37%, but the second lowest sensitivity at 55.30%.

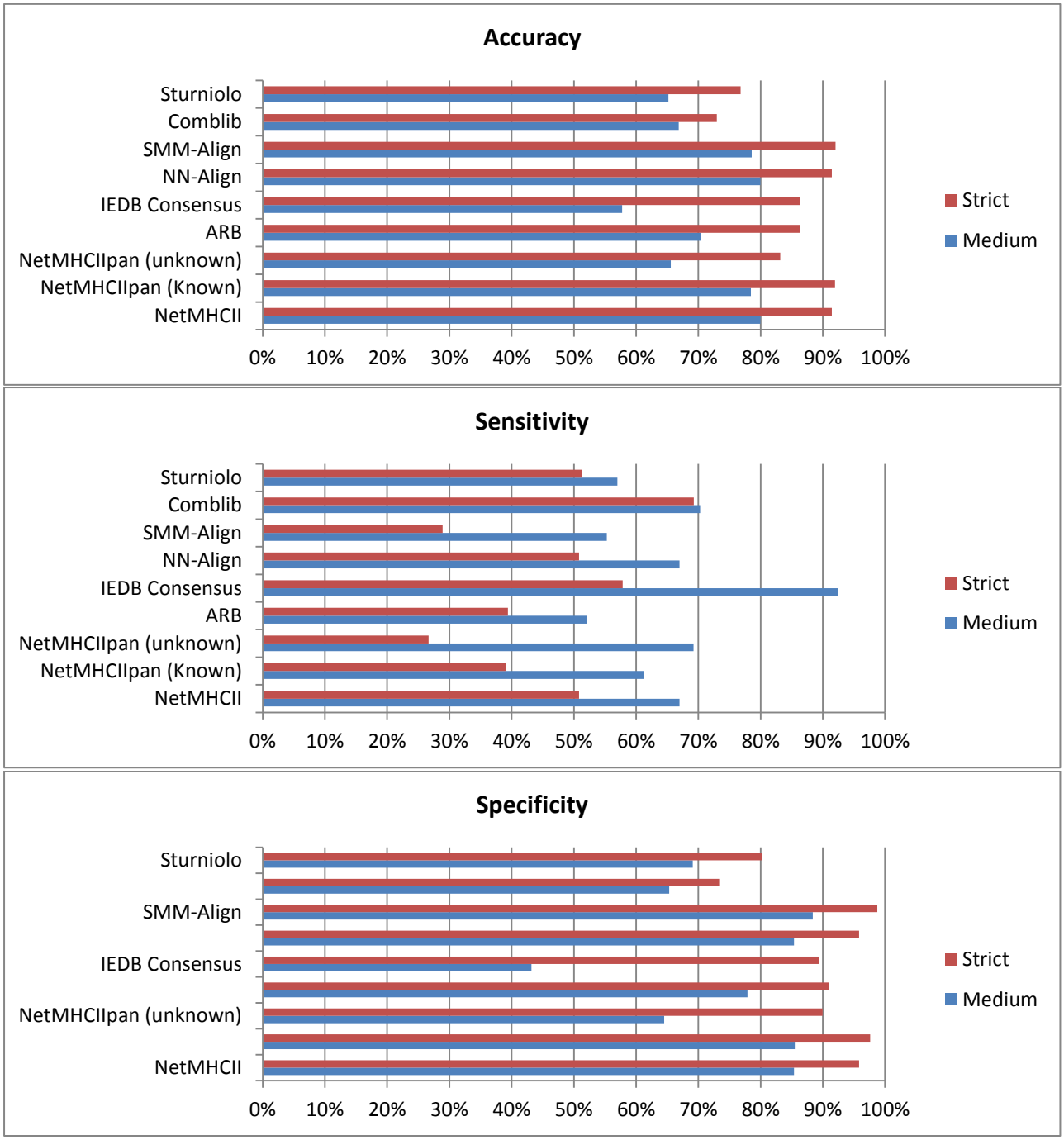


Figure 3.1: Comparison of Accuracy, Sensitivity and Specificity across epitope prediction tools. Results using strict and medium cut-off values are shown in red and blue respectively.

3.3.2 Results per HLA allele

Similarly to previous evaluation studies in literature, the accuracy of each tool differed by HLA allele. This is demonstrated in Figure 3.6 which shows the accuracy from each tool for 3 different HLA DRB1 alleles. Similar results were seen for all other HLA alleles.

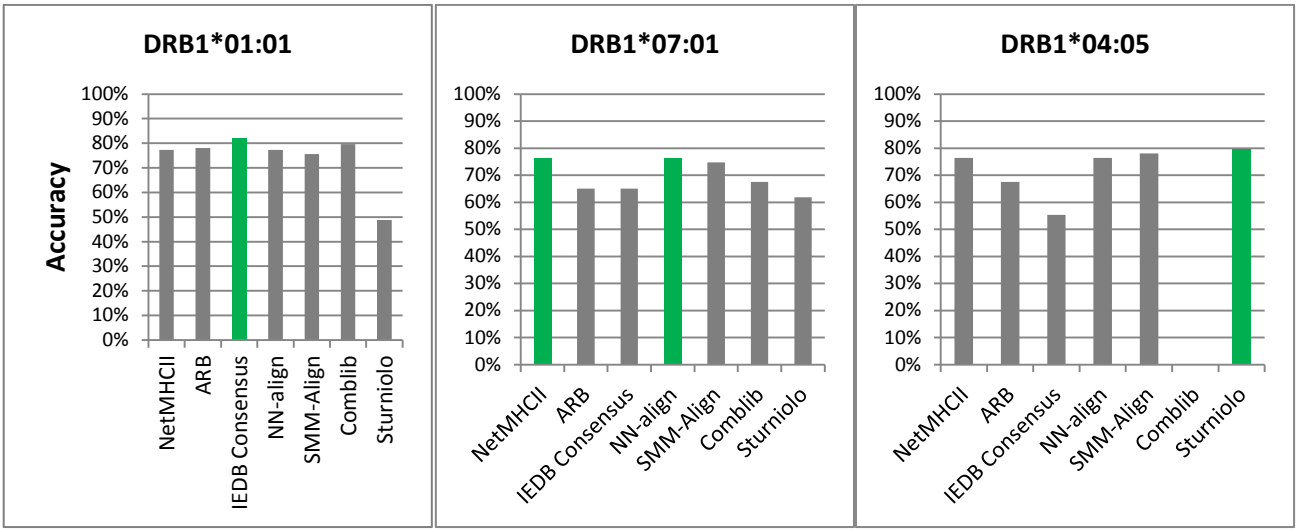


Figure 3.2: Accuracy of prediction tools by HLA allele. The prediction tool with the highest accuracy is shown in green, indicating there is no one overall most accurate tool to use but rather a most appropriate tool depending on which HLA allele is used in the prediction.

For each HLA allele, the tool with the highest accuracy, sensitivity, and specificity was determined. For those alleles that had at least 20 positive data points in the data when using the medium cut-off, the results were compared to the IEDB recommended tool (Table 3.5). The preferred tool to use was different depending on whether a high overall accuracy, sensitivity or specificity was preferred; highlighting the importance of determining which statistic is most preferable given the aim of the study. When considering accuracy and specificity, the best tool to use varies greatly by HLA allele; however when considering only the sensitivity of each allele, the IEDB consensus tool was the preferred tool for 16/19 alleles. For the majority of the alleles (18/19) the IEDB recommended tool is the IEDB consensus tool, possibly indicating a preference for higher sensitivity in prediction tools by the IEDB.

Table 3.5: Comparison of evaluation results with IEDB recommended tool by HLA allele. For each allele with at least 20 positive data points, the tool with the highest accuracy, sensitivity and specificity is shown. The IEDB recommended tool has also been shown.

Allele	Positive Data Points	Negative Data Points	Accuracy	Sensitivity	Specificity	IEDB Recommended
HLA-DQA1*01:02/DQB1*06:02	59	64	NetMHCIIpan	IEDB Consensus	ARB	Consensus
HLA-DQA1*03:01/DQB1*03:02	38	85	NetMHCII	IEDB Consensus	SMM-Align	Consensus
HLA-DQA1*04:01/DQB1*04:02	52	71	NetMHCII	IEDB Consensus	ARB	Consensus
HLA-DQA1*05:01/DQB1*02:01	53	70	IEDB Consensus	IEDB Consensus	ARB	Consensus
HLA-DQA1*05:01/DQB1*03:01	56	67	SMM-Align	NetMHCII	SMM-Align	Consensus
HLA-DRB1*01:01	75	48	NetMHCIIpan	ARB	Sturniolo	Consensus
HLA-DRB1*04:01	47	76	SMM-Align	IEDB Consensus	Sturniolo	Consensus
HLA-DRB1*04:04	39	84	NetMHCIIpan	IEDB Consensus	ARB	Consensus
HLA-DRB1*04:05	41	82	Sturniolo	IEDB Consensus	SMM-Align	Consensus
HLA-DRB1*07:01	39	84	NetMHCII	IEDB Consensus	SMM-Align	Consensus
HLA-DRB1*08:02	56	67	NetMHCII	IEDB Consensus	SMM-Align	Consensus
HLA-DRB1*09:01	58	65	NetMHCII	IEDB Consensus	SMM-Align	Consensus
HLA-DRB1*11:01	32	91	Sturniolo	IEDB Consensus	Sturniolo	Consensus
HLA-DRB1*12:01	24	99	SMM-Align	ARB	SMM-Align	SMM-align
HLA-DRB1*13:02	31	92	SMM-Align	IEDB Consensus	SMM-Align	Consensus
HLA-DRB1*15:01	25	98	NetMHCII	IEDB Consensus	SMM-Align	Consensus
HLA-DRB3*01:01	25	98	SMM-Align	IEDB Consensus	NetMHCIIpan	Consensus
HLA-DRB4*01:01	46	77	NetMHCII	IEDB Consensus	SMM-Align	Consensus
HLA-DRB5*01:01	33	90	NetMHCIIpan	IEDB Consensus	NetMHCII	Consensus

3.4 Conclusion

Given the wide range of options available to researchers when choosing an epitope prediction tool, it is important to evaluate the accuracy of these tools in the context of the purpose for which the investigation is carried out. In the current project, *M. tuberculosis* proteins are being investigated in order to identify epitopes that may be possible vaccine candidates.

Epitope prediction tools are most commonly used when searching for possible vaccine targets within reverse vaccinology pipelines. Resulting epitopes would be fed into wet lab experiments for validation. Therefore researchers need to weigh up the advantage of decreasing the number of epitopes to be fed into validation from a time and cost perspective (i.e. increasing specificity), but at the same time limiting the number of true positive epitopes missed (i.e. increasing sensitivity). A perfect tool would have high values for both sensitivity and specificity but this was not the case with the eight epitope prediction tools tested above. Tools with a high sensitivity had a low specificity and vice versa, therefore necessitating the need to choose a preference for one over the other. When comparing the most appropriate tool for each HLA allele to the IEDB recommended tool, in most cases the tool with the highest sensitivity matched the IEDB recommended tool, indicating a possible preference for higher sensitivity by the IEDB. This is reasonable given the aim of identifying vaccine candidates and therefore limiting the number of true positive results that may be missed.

Experimental data used within this evaluation were derived from *M. tuberculosis* antigens only, compared to previously reported valuations which used antigens from many source pathogens. This was done in order to determine whether a specific tool may perform better for *M. tuberculosis* proteins. Based on a preference for sensitivity, the IEDB consensus tool was the preferred tool for the majority of the alleles included in the evaluation study. This corresponded to results obtained by the IEDB when using a large number (>10,000) of assay results from a variety of pathogens (peptides for 114 proteins from 30 organisms) (Wang *et al*. 2008). Therefore no difference in the preferred tool was found specifically for *M. tuberculosis* proteins. Even though the *M. tuberculosis* proteome may have unique characteristics from other bacteria and pathogens, the short length of the majority of MHC class II epitopes (15 mers) may result in no significant differences in the structure of T-cell epitopes between pathogens.

There was a high concordance between the tool with the best sensitivity for each allele according to the analysis within this chapter and the IEDB recommended tool for each allele (16/19 alleles). For 3/19 alleles, the recommended tool from this study based on sensitivity of the tools did not match the IEDB recommended tool. However, given that no significant difference was found when using *M. tuberculosis* proteins only, and that the evaluation by the IEDB was performed using a significantly larger amount of data leading to more reliable results, the IEDB recommended tool will also be used for these three alleles. A larger number of alleles than what was used within this evaluation will be used in the prediction of epitopes within the PPE_MPTR proteins (Chapter 4). The IEDB recommended tool for all alleles will be used for the remainder of this project.

The results of this analysis show that no tool is 100% accurate, and therefore the limitations of these tools should be kept in mind when analysing the results from epitope predictions, and viewed in the context of the study goal. For this study, an overall sensitivity of 92.49% for the IEDB consensus tool is considered to be suitably high for the aim of identifying novel vaccine candidates when used within a reverse vaccinology workflow.

3.5 References

Bui, H. *et al*., 2005. Automated generation and evaluation of specific MHC binding predictive tools: ARB matrix applications. *Immunogenetics*, 57(5), pp.304–14.

El-Manzalawy, Y., Dobbs, D. & Honavar, V., 2008. On evaluating MHC-II binding peptide prediction methods. *PloS one*, 3(9), p.e3268.

Karosiene, E. *et al*., 2013. NetMHCIIpan-3.0, a common pan-specific MHC class II prediction method including all three human MHC class II isotypes, HLA-DR, HLA-DP and HLA-DQ. *Immunogenetics*, 65(10), pp.711–24.

Kim, Y. *et al*., 2012. Immune epitope database analysis resource. *Nucleic acids research*, 40(Web Server issue), pp.W525–30.

Lin, H.H. *et al*., 2008. Evaluation of MHC-II peptide binding prediction servers: applications for vaccine research. *BMC bioinformatics*, 9 Suppl 12, p.S22.

Lundegaard, C. *et al*., 2008. NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8-11. *Nucleic Acids Research*, 36(Web Server), pp.W509–W512.

Nielsen, M. & Lund, O., 2009. NN-align. An artificial neural network-based alignment algorithm for MHC class II peptide binding prediction. *BMC bioinformatics*, 10, p.296.

Nielsen, M., Lundegaard, C. & Lund, O., 2007. Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method. *BMC bioinformatics*, 8, p.238.

Singh, H. & Raghava, G.P., 2001. ProPred: prediction of HLA-DR binding sites. *Bioinformatics (Oxford, England)*, 17(12), pp.1236–7.

Sturniolo, T. *et al*., 1999. Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices. *Nature biotechnology*, 17(6), pp.555–61.

Wang, P. *et al*., 2008. A systematic assessment of MHC class II peptide binding predictions and evaluation of a consensus approach. *PLoS computational biology*, 4(4), p.e1000048.

Wang, P. *et al*., 2010. Peptide binding predictions for HLA DR, DP and DQ molecules. *BMC bioinformatics*, 11(1), p.568.

Zhang, L. *et al*., 2012. TEPITOPEpan: extending TEPITOPE for peptide binding prediction covering over 700 HLA-DR molecules. *PloS one*, 7(2), p.e30483.

Chapter 4: MHC Class II Epitope Prediction

4.1 Introduction

Given previous evidence of the interaction between the PE/PPE proteins with the host immune system, including eliciting CD4+ and CD8+ T-cell responses (Sampson 2011) (Chapter 1, section 1.1.6), it is hypothesized that multiple epitopes, defined as short peptides capable of producing an immune response, exist within the *Mycobacterium tuberculosis* PPE_MPTR proteins. The PPE_MPTR proteins consist of a conserved PPE region at the start of the encoded protein and a variable MPTR region which may contain differential epitope density.

Given the large number of PPE_MPTR proteins, with certain members greater than 3000 amino acids, and the large number of possible HLA alleles in the human population, the number of experiments that would be needed in conventional wet lab approaches used to ascertain host-pathogen interactions would be restrictively high. Epitope prediction which makes use of various *in silico* methods can offer a solution. Using the results from an *in silico* approach the number of experiments needed in subsequent validation steps can be reduced, resulting in both cost and time savings.

Based on the evaluation of 8 well known epitope prediction tools using known *M. tuberculosis* epitopes (Chapter 3), the IEDB recommended tool for each HLA allele has been used to predict epitopes within the *M. tuberculosis* PPE_MPTR proteins.

4.2 Methods

4.2.1 PPE_MPTR protein sequence data

The protein sequence for each of the 23 *M. tuberculosis* PPE_MPTR proteins for the reference strain H37Rv was downloaded from TubercuList (Lew *et al*. 2011). A multiple sequence alignment of all 23 PPE_MPTR protein sequences was performed using Clustal Omega (Sievers *et al*. 2011), using default parameter values. Based on this alignment, three PPE_MPTR proteins were not included in the analysis since the PPE region in these proteins was either not similar in size or sequence to the other PPE_MPTR proteins (Rv0304c and Rv0354c), or completely missing (Rv2353). Therefore a total of 20 PPE_MPTR proteins were included in the analysis (Table 4.1).

Table 4.1: PPE_MPTR proteins. The size of the DNA (bp) and protein (aa) sequence for the 20 PPE_MPTR proteins included in the analysis is shown. (Sampson 2011).

PE/PPE Number	Rv number	bp size	aa size
PPE_MPTR6	Rv0305c	2892	964
PPE_MPTR8	Rv0355c	9903	3301
PPE_MPTR10	Rv0442c	1464	488
PPE_MPTR12	Rv0755c	1938	646
PPE_MPTR13	Rv0878c	1332	444
PPE_MPTR16	Rv1135c	1857	619
PPE_MPTR21	Rv1548c	2037	679
PPE_MPTR24	Rv1753c	3162	1054
PPE_MPTR28	Rv1800	1968	656
PPE_MPTR34	Rv1917c	4380	1460
PPE_MPTR35	Rv1918c	2964	988
PPE_MPTR40	Rv2356c	1848	616
PPE_MPTR42	Rv2608	1743	581
PPE_MPTR52	Rv3144c	1230	410
PPE_MPTR53	Rv3159c	1773	591
PPE_MPTR54	Rv3343c	7572	2524
PPE_MPTR55	Rv3347c	9474	3158
PPE_MPTR56	Rv3350c	11151	3717
PPE_MPTR62	Rv3533c	1749	583
PPE_MPTR64	Rv3558	1659	553

4.2.2 HLA alleles

The most common HLA class II alleles in the general human population were included in the analysis (Greenbaum *et al*. 2011). This list of 27 alleles represent >99% population coverage, based on data available from DbMHC (Helmberg *et al*. 2004) and allelefrequencies.net (González-Galarza *et al*. 2015). This list is used within the IEDB analysis resource (Kim *et al*. 2012). In addition, to ensure relevance to study populations most affected by TB, the most frequent DRB1 alleles in the 22 high burden countries specified by WHO (pre-2015) (WHO 2014), that were not part of the original list by Greenbaum *et al*. (2011) were also included in the analysis. For each of the 22 high burden countries, DRB1 allele frequencies were obtained from allelefrequencies.net (González-Galarza *et al*. 2015), and those with a greater than 5% frequency were included. In total, 47 HLA class II alleles were included in the analysis (Table 4.2).

4.2.3 Prediction of epitopes

Each of the 20 PPE_MPTR protein sequences was segmented into overlapping peptides of 15 amino acids. Epitope prediction for each of the overlapping peptides was performed against each of the 47 selected HLA class II alleles. Based on the evaluation of MHC class II prediction tools (Chapter 3), the IEDB recommended tool for each of the alleles was used, and is shown in Table 4.2.

Table 4.2: HLA alleles and the IEDB recommended tool for each allele. Alleles in the original list by Greenbaum *et al*. 2011 are indicated by (*)

	Allele	IEDB Recommended Tool		Allele	IEDB Recommended Tool
1.	*HLA-DPA1*01:01-DPB1*04:01*(*)	IEDB Consensus	**25.**	*HLA-DRB1*10:01*	NetMHCIIpan
2.	*HLA-DPA1*01:03-DPB1*02:01*(*)	IEDB Consensus	**26.**	*HLA-DRB1*11:01*(*)	IEDB Consensus
3.	*HLA-DPA1*02:01-DPB1*01:01*(*)	IEDB Consensus	**27.**	*HLA-DRB1*11:02*	Sturniolo
4.	*HLA-DPA1*02:01-DPB1*05:01*(*)	IEDB Consensus	**28.**	*HLA-DRB1*11:03*	NetMHCIIpan
5.	*HLA-DPA1*03:01-DPB1*04:02*(*)	IEDB Consensus	**29.**	*HLA-DRB1*11:04*	Sturniolo
6.	*HLA-DQA1*01:01-DQB1*05:01*(*)	IEDB Consensus	**30.**	*HLA-DRB1*12:01*(*)	IEDB Consensus
7.	*HLA-DQA1*01:02-DQB1*06:02*(*)	IEDB Consensus	**31.**	*HLA-DRB1*12:02*	NetMHCIIpan
8.	*HLA-DQA1*03:01-DQB1*03:02*(*)	IEDB Consensus	**32.**	*HLA-DRB1*13:01*	Sturniolo
9.	*HLA-DQA1*04:01-DQB1*04:02*(*)	IEDB Consensus	**33.**	*HLA-DRB1*13:02*(*)	IEDB Consensus
10.	*HLA-DQA1*05:01-DQB1*02:01*(*)	IEDB Consensus	**34.**	*HLA-DRB1*13:03*	IEDB Consensus
11.	*HLA-DQA1*05:01-DQB1*03:01*(*)	IEDB Consensus	**35.**	*HLA-DRB1*14:01*	IEDB Consensus
12.	*HLA-DRB1*01:01*(*)	IEDB Consensus	**36.**	*HLA-DRB1*14:04*	IEDB Consensus
13.	*HLA-DRB1*01:02*	Sturniolo	**37.**	*HLA-DRB1*14:05*	IEDB Consensus
14.	*HLA-DRB1*03:01*(*)	IEDB Consensus	**38.**	*HLA-DRB1*15:01*(*)	IEDB Consensus
15.	*HLA-DRB1*03:02*	IEDB Consensus	**39.**	*HLA-DRB1*15:02*	Sturniolo
16.	*HLA-DRB1*04:01*(*)	IEDB Consensus	**40.**	*HLA-DRB1*15:03*	NetMHCIIpan
17.	*HLA-DRB1*04:03*	NetMHCIIpan	**41.**	*HLA-DRB1*15:04*	IEDB Consensus
18.	*HLA-DRB1*04:05*(*)	IEDB Consensus	**42.**	*HLA-DRB1*16:01*	IEDB Consensus
19.	*HLA-DRB1*04:06*	NetMHCIIpan	**43.**	*HLA-DRB1*16:02*	IEDB Consensus
20.	*HLA-DRB1*07:01*(*)	IEDB Consensus	**44.**	*HLA-DRB3*01:01*(*)	IEDB Consensus
21.	*HLA-DRB1*08:02*(*)	IEDB Consensus	**45.**	*HLA-DRB3*02:02*(*)	NetMHCIIpan
22.	*HLA-DRB1*08:03*	IEDB Consensus	**46.**	*HLA-DRB4*01:01*(*)	IEDB Consensus
23.	*HLA-DRB1*08:04*	Sturniolo	**47.**	*HLA-DRB5*01:01*(*)	IEDB Consensus
24.	*HLA-DRB1*09:01*(*)	IEDB Consensus			

The collection of IEDB recommended tools was downloaded from the IEDB analysis resource in the form of a mixture of python scripts and Linux 32-bit environment specific binaries, and run locally through command line operations (Kim *et al*. 2012). When choosing the IEDB recommended option when running the epitope prediction, prediction results are in the form of a percentile rank, which is found by comparing a peptide's binding score against the scores of five million random 15 mers selected from the SWISSPROT database (Kim *et al*. 2012). A low percentile rank indicates high binding affinity compared to other possible peptides. Peptides with a percentile rank of <1% for at least one HLA allele were classified as binders and therefore as predicted epitopes.

4.2.4 Determination of the PPE boundary

Given the hypothesis that genetic diversity within PPE_MPTR proteins may differentially modulate human immune response, and that the PPE region is known to be conserved compared to the variable MPTR region within the PPE_MPTR proteins, it was important to determine the boundary between the two regions. The PPE region is between 170-180 amino acids long and highly conserved between different strains (Cole *et al*. 1998), but also relatively conserved between the PPE_MPTR proteins themselves. The amino acid sequence of each of the 20 PPE_MPTR proteins from H37Rv were aligned using Clustal Omega (Sievers *et al*. 2011), using default parameters, and this alignment was visually inspected to determine the end of the conserved PPE region and the start of the variable MPTR region. The number of amino acids from the start of the protein until this boundary for each of the PPE_MPTR proteins was counted (results were between 172 – 176 amino acids). This boundary was used to determine and compare the number of predicted epitopes within the PPE region versus the MPTR region for each protein.

In an analysis of the epitope density within the PE_PGRS proteins, epitope density fluctuated along the length of each protein (Copin *et al*. 2014). Therefore, in addition to comparing epitope density within the PPE versus the MPTR region, each protein was segmented into sections of 60 amino acids along the length of the protein and the number of epitopes in each segment counted.

4.2.5 Identification of promiscuous epitopes

A peptide is classified as a binder and therefore a predicted epitope if it is able to bind to at least one HLA allele (percentile rank < 1%). A peptide may be able to bind to more than one HLA allele. For each peptide classified as a binder, the number of HLA alleles (out of the 47 alleles included) that the peptide is predicted to bind to was counted. Peptides that were predicted to bind to more than one allele were classified as promiscuous epitopes. The percentage of promiscuous epitopes was compared within the PPE region versus the MPTR region as well as along the length of each protein (in segments of 60 amino acids).

4.2.6 Binding ability of HLA alleles

Literature has shown that specific alleles may be able to bind more universally than others (i.e. able to bind to more peptides than others) (Paul *et al*. 2013). In order to test whether this is true for the PPE_MPTR proteins and the 47 alleles used in the prediction, the number of peptides that each of the 47 HLA alleles was able to bind to was counted and compared. This was investigated overall for the PPE_MPTR proteins as well as for each of the proteins individually. The frequency of the high binding HLA alleles within different populations is an important consideration for the potential population coverage of an epitope when used in a vaccine cocktail. This is investigated in Chapter 6.

4.3 Results

4.3.1 Number of epitopes in PPE versus MPTR region

A total number of 4,548 epitopes were predicted across the 20 PPE_MPTR proteins. Figure 4.1 shows the density of predicted epitopes for each of the 20 proteins, in the PPE region versus the MPTR region. Density of epitopes is calculated as the number of epitopes in a section weighted by the length of that section.

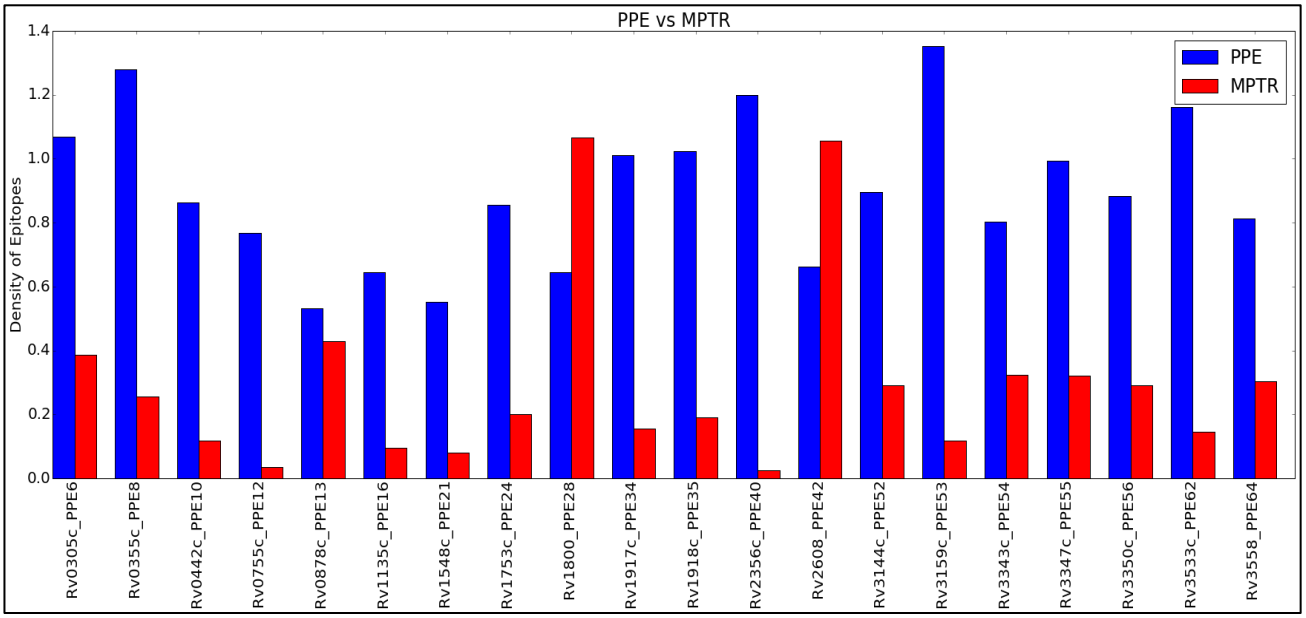


Figure 4.1: Density of epitopes in PPE vs. MPTR region. The density of epitopes in the PPE and MPTR region of the protein is shown in blue and red respectively. Density is calculated as the number of predicted epitopes weighted by the length of each of the respective sections.

There is a higher density of predicted epitopes in the PPE versus the MPTR region of each protein (Figure 4.1). A variable density of epitopes in the MPTR region can be seen across the different proteins, with certain proteins having almost no epitopes in the MPTR region compared to the PPE region (Rv0755c, Rv2356c), while others show a high density of epitopes comparable with the PPE region (Rv1800, Rv2608).

Mean epitope density in the PPE region is on average 3-fold higher than the epitope density in the MPTR region across all 20 PPE_MPTR proteins. This is statistically significantly different (P< 0.001) (Figure 4.2).

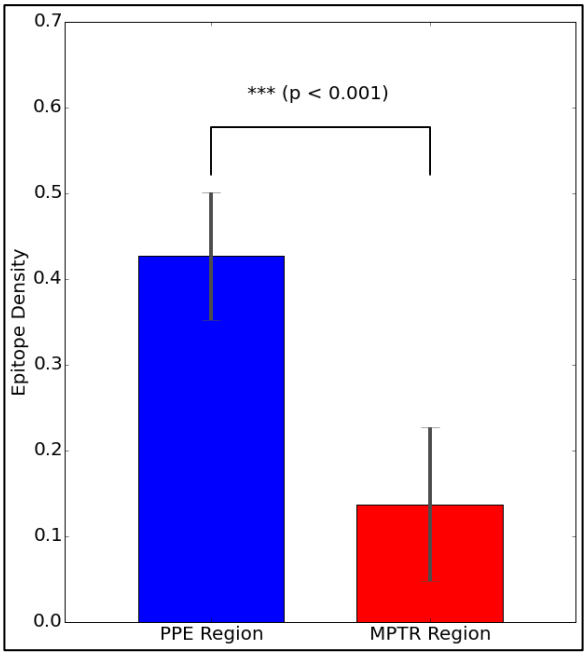


Figure 4.2: Mean density of epitopes in PPE region versus MPTR region. The PPE region had an overall epitope density of 0.43 compared to an overall epitope density of 0.14 in the MPTR region across all PPE_MPTR proteins. This difference is statistically significantly different.

4.3.2 Patterns of fluctuation along the length of the protein

When comparing the number of epitopes in each 60 amino acid segment along the length of each protein, three major patterns of epitope distribution can be seen (Table 4.3).

Table 4.3: Observed patterns of fluctuation. Three patterns of fluctuation were identified when investigating epitope density along the length of the protein for each of the PPE_MPTR proteins.

Pattern Observed	Description
Fluctuating	Alternating areas of high and low number of epitopes can be seen along the entire length of the protein (Rv0305c, Rv3055c, Rv0878c, Rv1753c, Rv1917c, Rv1918c, Rv3144c, Rv3343c, Rv3347c, and Rv3350c).
Decreasing	A high number of epitopes in the PPE region is seen which decreases substantially in the MPTR region (Rv0442c, Rv0755c, Rv1135c, Rv1548c, Rv2356c, Rv3159c, Rv3533c, and Rv3556).
Increasing/Stable	The number of epitopes either increases or is stable in the PPE region versus the MPTR region (Rv1800 and Rv2608).

Figure 4.3 below shows an example of each of the patterns observed.

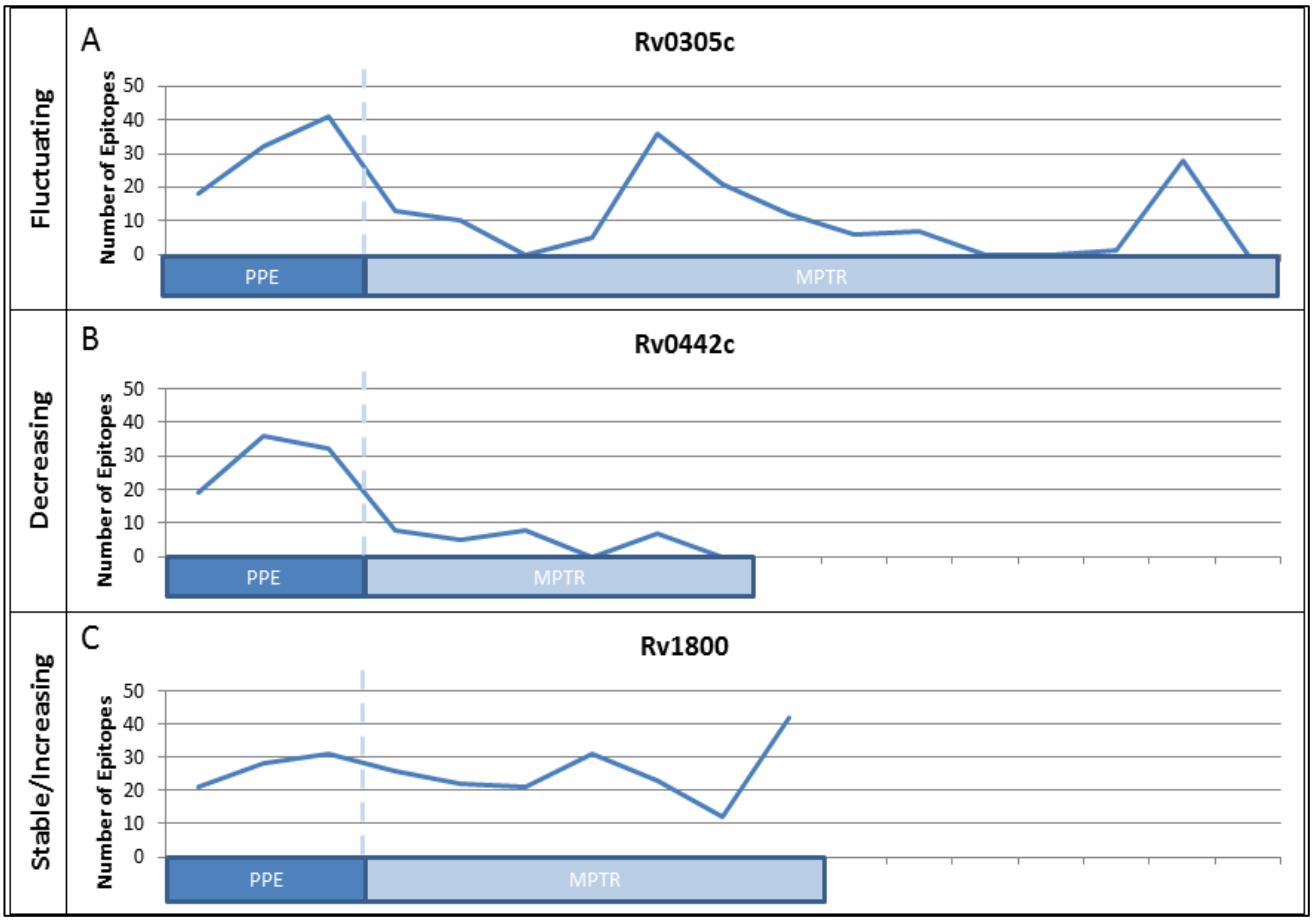


Figure 4.3: Patterns of fluctuation observed in the number of predicted epitopes along the length of PPE_MPTR proteins. Three patterns of fluctuation were seen along the length of the PPE_MPTR proteins. An example of each has been shown.

The density of predicted epitopes along the length of the protein for all of the PPE_MPTR proteins can be found in Appendix B, with the pattern of fluctuation (fluctuating, decreasing or increasing/stable) given for each protein.

4.3.3 Promiscuous epitopes

A given epitope will elicit an immune response only in individuals expressing HLA molecules capable of binding to that particular epitope. The higher the number of HLA alleles an epitope is able to bind to, the higher the potential population coverage when the peptide is included in a peptide based vaccine. To explore this in the context of the PPE_MPTR proteins, the number of HLA alleles that each predicted epitope is able to bind to was counted and the distribution compared between the PPE region and MPTR region (Figure 4.4).

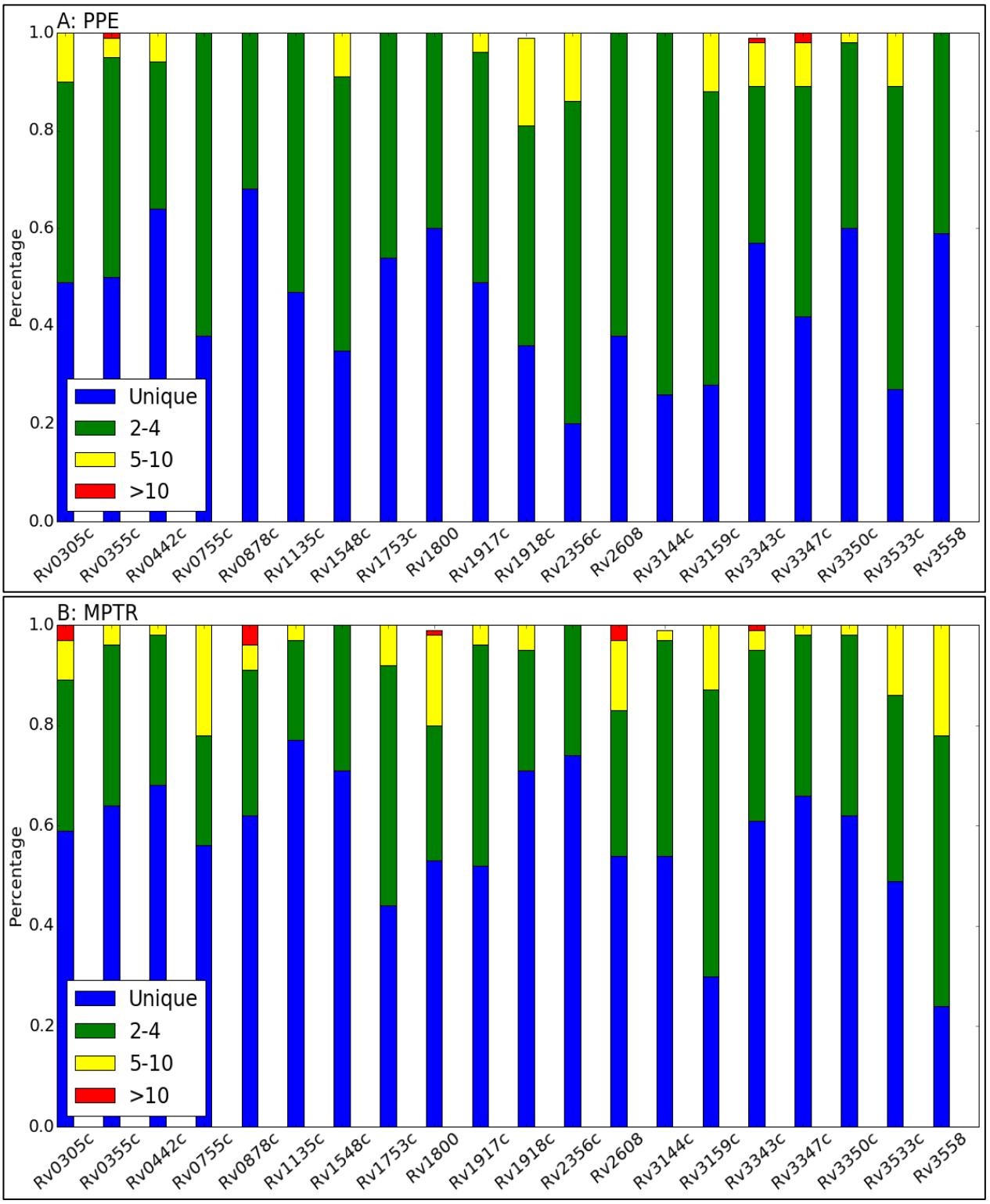


Figure 4.4: Distribution of promiscuous epitopes. Blue bars indicate the percentage of peptides that are only predicted to bind to one unique HLA allele. Green, yellow and red bars represent the number of peptides that are able to bind to more than one HLA allele (promiscuous epitopes). This is shown separately for the PPE region and MPTR region in A) and B) respectively.

Overall, the majority of predicted epitopes (56%) bind to only one unique HLA allele. A further 22% are able to bind to only 2 HLA alleles, and 11% to only 3 HLA alleles. There is therefore a high peptide to HLA allele binding specificity, with less than 1% of peptides able to bind to 10 or more HLA alleles. The maximum number of HLA alleles that any one peptide was predicted to bind to was 27 (representing 0.04% of the total number of predicted epitopes). These peptides come from Rv2608, which also has the second highest percentage of peptides predicted to bind to 10 or more HLA alleles (after Rv0878c). Rv2608 is part of the ID93 subunit vaccine candidate currently undergoing clinical trials.

The distribution in the number of HLA alleles per peptide varies by protein, with certain proteins having a larger percentage of promiscuous epitopes than others. Other PPE_MPTR proteins with a high percentage of promiscuous epitopes include Rv3558 (which contains the lowest percentage of unique binders), Rv0305c, Rv0355c, Rv0878c, Rv1800, Rv3343c, Rv3347c and Rv3350c which all contain epitopes able to bind to 10 or more alleles.

The overall percentage of promiscuous epitopes in the PPE region and the MPTR region was compared to determine whether characteristics within the respective regions may be contributing to promiscuity (Figure 4.5). A significant difference in the proportion of promiscuous epitopes within the PPE region versus the MPTR region across all PPE_MPTR proteins was found ($p < 0.01$).

The percentage of binders versus non-binders, unique binders versus promiscuous binders, and distribution of promiscuous epitopes for each PPE_MPTR protein is shown in Appendix A.

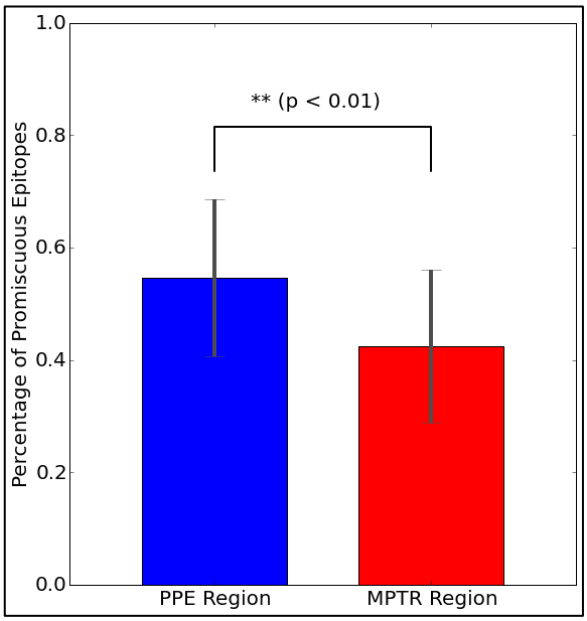


Figure 4.5: Percentage of promiscuous epitopes. Percentage of promiscuous epitopes in the PPE region (blue) compared to the MPTR region (red) is statistically significantly different ($p < 0.01$).

The distribution of promiscuous epitopes along the length of each protein (in segments of 60 amino acids) was investigated to determine if promiscuous epitopes localise to certain regions. Results for Rv2608 and Rv3558 are shown in Figure 4.6. Results for all PPE_MPTR proteins can be found in Appendix B. No apparent correlation between the proportion of promiscuous epitopes and location along the length of the protein was seen other than when comparing PPE region versus MPTR region.

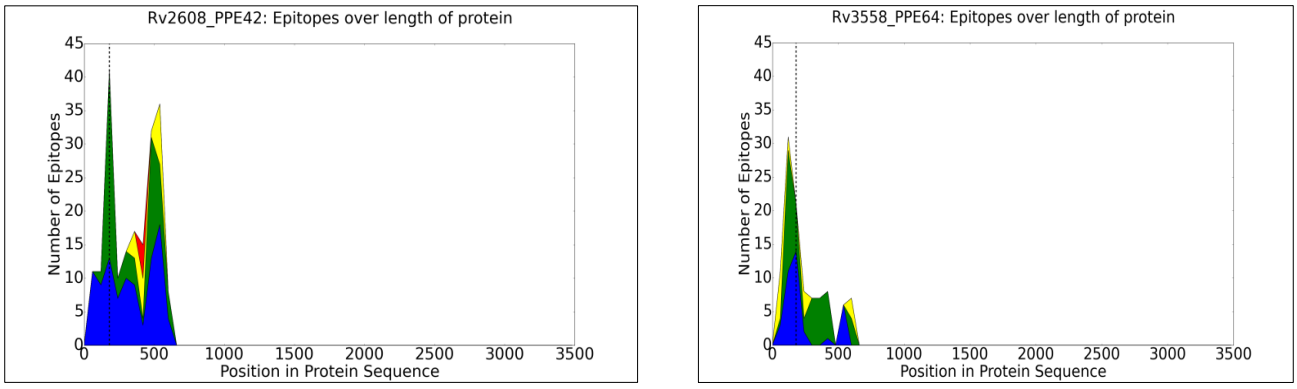


Figure 4.6: Distribution of promiscuous epitopes along the length of the protein. Colours consistent with Figure 4.4 show where promiscuous epitopes are predicted along the length of the protein.

4.3.4 Binding ability of HLA alleles

Figure 4.5 shows the percentage of peptides each of the 47 HLA alleles are able to bind to, overall for all of the PPE_MPTR proteins.

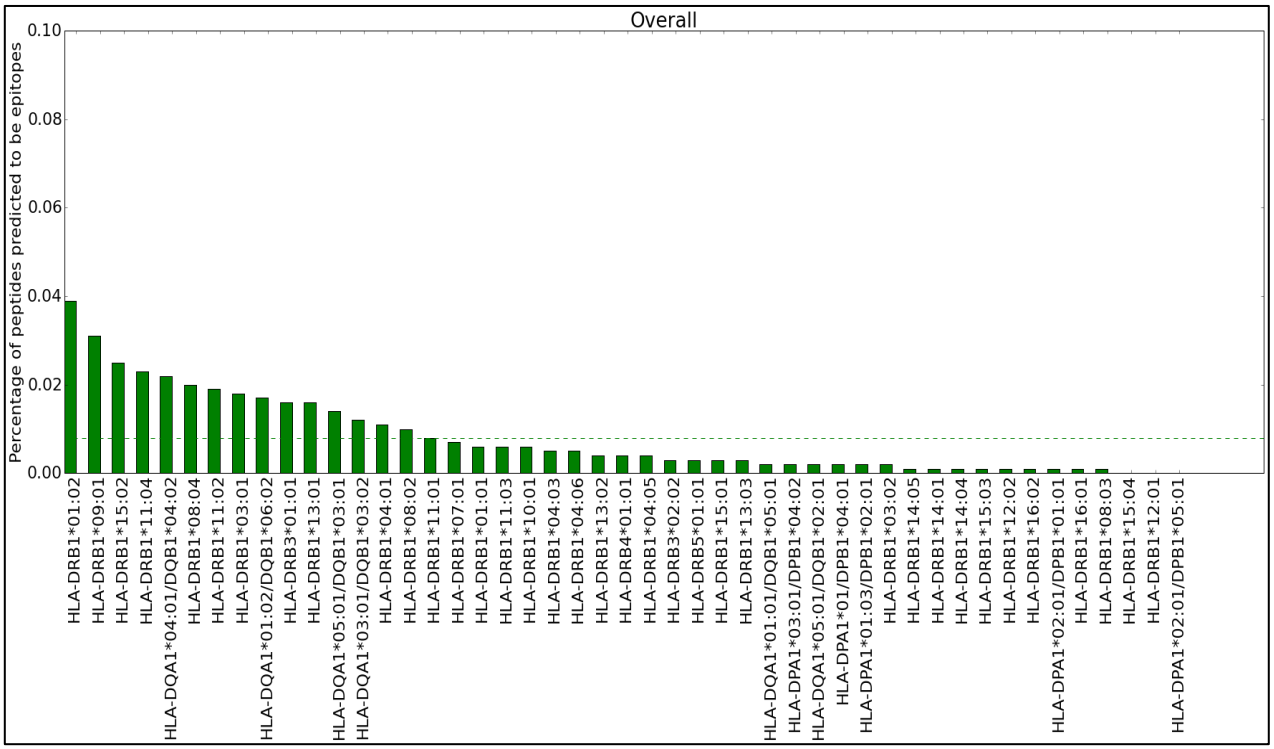


Figure 4.7: Binding ability of HLA alleles. Percentage of peptides each HLA class II allele is able to bind to is shown above.

It is evident that certain HLA class II alleles are more universal binders than others. The percentage of peptides each of the HLA alleles are able to bind to was also investigated for each of the PPE_MPTR proteins individually. These results are shown in Appendix C. The distribution changes significantly when comparing different PPE_MPTR proteins. For example, when all of the PPE_MPTR proteins are considered together, *HLA-DRB1*01:02* is able to bind to the highest number of peptides, however when looking specifically at Rv0442c, *HLA-DQA1*04:01-DQB1*04:02* is able to bind to the highest number of peptides. It is also well known, that HLA alleles are found in different frequencies in different populations (Chapter 2). Therefore depending on which PPE_MPTR protein is used within a vaccine cocktail, the population coverage will change depending on the frequency of the highest binding alleles in the population of interest. This is an important consideration when choosing vaccine candidates, and could be used to customise vaccine regimes for specific populations. This is considered in more detail in Chapter 6.

4.4 Conclusion

A large number of epitopes are predicted within the PPE_MPTR proteins, supporting the hypothesis that the PPE_MPTR proteins may play a role in host pathogen interactions. A larger density of epitopes is found within the conserved PPE region, while the density of epitopes within the MPTR region varies by protein. The PPE region is thought to be highly conserved and therefore a higher density of predicted epitopes in this region is consistent with literature showing conservation of *M. tuberculosis* epitopes (Comas *et al*. 2010). Fluctuating patterns of epitopes have been identified along the length of the PPE_MPTR proteins. Possible reasons for the observed fluctuating patterns may include:

- Areas of conservation versus diversity. The PPE_MPTR proteins are highly variable, and are considered to be hotspots for recombination, while *M. tuberculosis* epitopes have been shown to be conserved across strains. Therefore the fluctuating patterns of epitopes along the length of the protein may be due to a large number of epitopes in regions of genetic conservation followed by low numbers in regions of diversity. However as hypothesized, genetically variable regions may play a role in antigenic variation, and therefore fluctuating patterns of epitopes may be due to a large number of epitopes in regions of genetic diversity rather than conservation. The MPTR section of the protein is characterised by imperfect repeat regions which also may contribute to the presence or lack of epitopes within these regions. Certain members of the PPE_MPTR family are known to contain large insertions and deletions adding to the genetic complexity of these proteins which may also contribute to the presence or lack of epitopes. The effect of genetic variation on epitope density will be further explored in Chapter 5.

- Areas of the protein that are more (or less) accessible to the host immune system due to the 3D protein structure may contain more (or less) epitopes. The 3D protein structure of the PPE_MPTR proteins is largely unknown (Strong *et al*. 2006), due to the difficulty in solubly expressing, isolating and crystallizing these proteins, and therefore this hypothesis is difficult to confirm. However, CD4+ T-cell epitopes are derived from *M. tuberculosis* proteins that have been phagocytised by human immune cells such as macrophages and dendritic cells and degraded within phagosomes before being presented by HLA class II molecules on the surface of the cell. The 3D structure of internalised proteins should therefore not play an important role in whether or not T-cell epitopes are found within these regions.

While most predicted epitopes show a high peptide to HLA allele binding specificity, there are certain peptides that are able to bind to 10 or more HLA alleles, therefore increasing the possible population coverage of any potential vaccine containing these peptides. Certain alleles that may be more or less frequent in different populations are also able to bind to a higher proportion of peptides than others. This may impact on protective efficacy of PPE-based subunit vaccines in different populations. Predicted population coverage and possible vaccine candidates are further explored in Chapter 6.

4.5 References

Cole, S.T. *et al*., 1998. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature*, 393(6685), pp.537–44.

Comas, I. *et al*., 2010. Human T cell epitopes of *Mycobacterium tuberculosis* are evolutionarily hyperconserved. *Nature genetics*, 42(6), pp.498–503.

Copin, R. *et al*., 2014. Sequence diversity in the pe_pgrs genes of *Mycobacterium tuberculosis* is independent of human T cell recognition. *mBio*, 5(1), pp.e00960–13.

González-Galarza, F.F. *et al*., 2015. Allele frequency net 2015 update: new features for HLA epitopes, KIR and disease and HLA adverse drug reaction associations. *Nucleic acids research*, 43(Database issue), pp.D784–8.

Greenbaum, J. *et al*., 2011. Functional classification of class II human leukocyte antigen (HLA) molecules reveals seven different supertypes and a surprising degree of repertoire sharing across supertypes. *Immunogenetics*, 63(6), pp.325–35.

Helmberg, W., Dunivin, R. & Feolo, M., 2004. The sequencing-based typing tool of dbMHC: typing highly polymorphic gene sequences. *Nucleic acids research*, 32(Web Server issue), pp.W173–5.

Kim, Y. *et al*., 2012. Immune epitope database analysis resource. *Nucleic acids research*, 40(Web Server issue), pp.W525–30.

Lew, J.M. *et al*., 2011. TubercuList--10 years after. *Tuberculosis (Edinburgh, Scotland)*, 91(1), pp.1–7.

Paul, S. *et al*., 2013. HLA class I alleles are associated with peptide-binding repertoires of different size, affinity, and immunogenicity. *Journal of immunology (Baltimore, Md. : 1950)*, 191(12), pp.5831–9.

Sampson, S.L., 2011. Mycobacterial PE/PPE proteins at the host-pathogen interface. *Clinical and Developmental Immunology*, 2011(Figure 1).

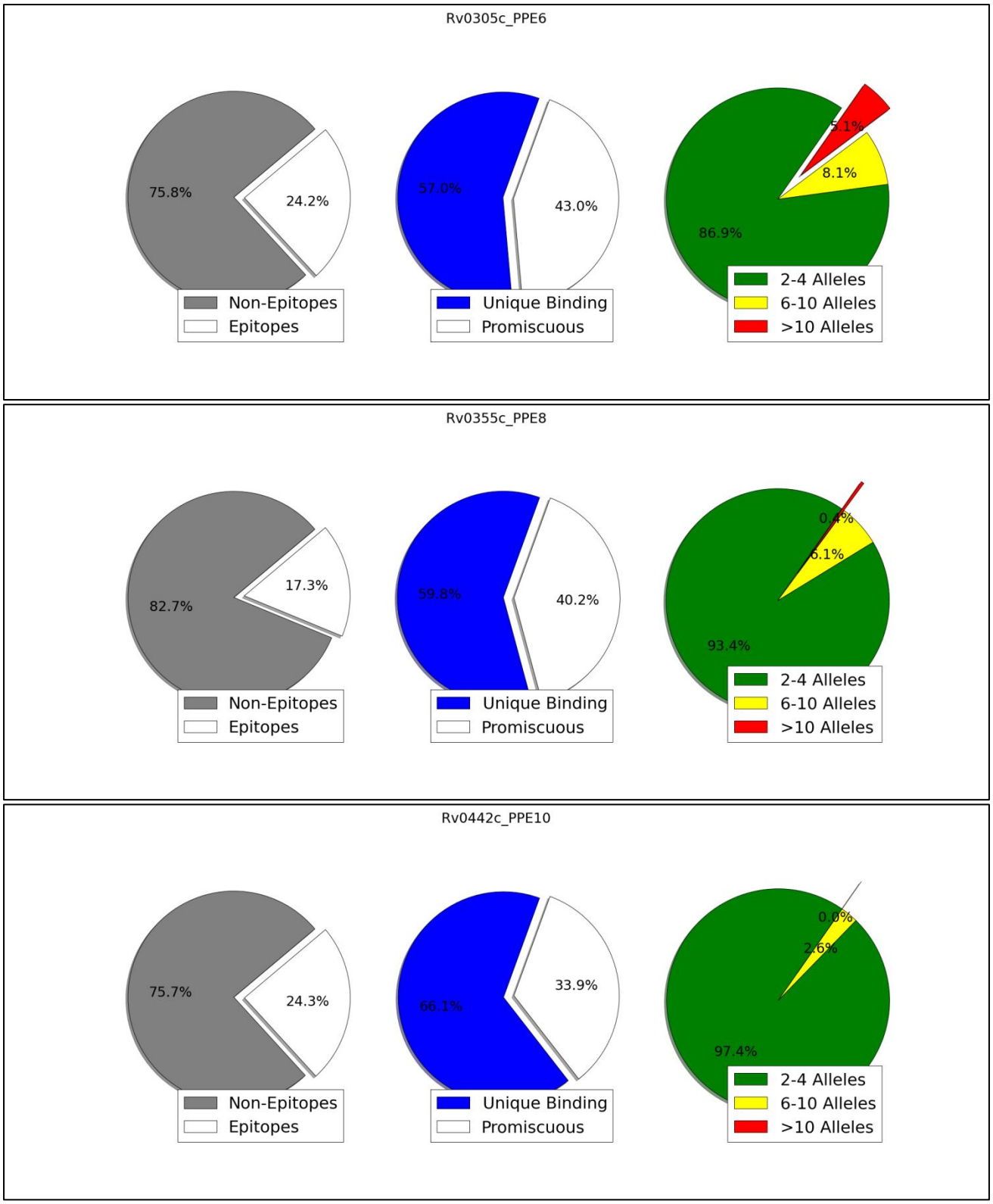
Sievers, F. *et al*., 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular systems biology*, 7, p.539.

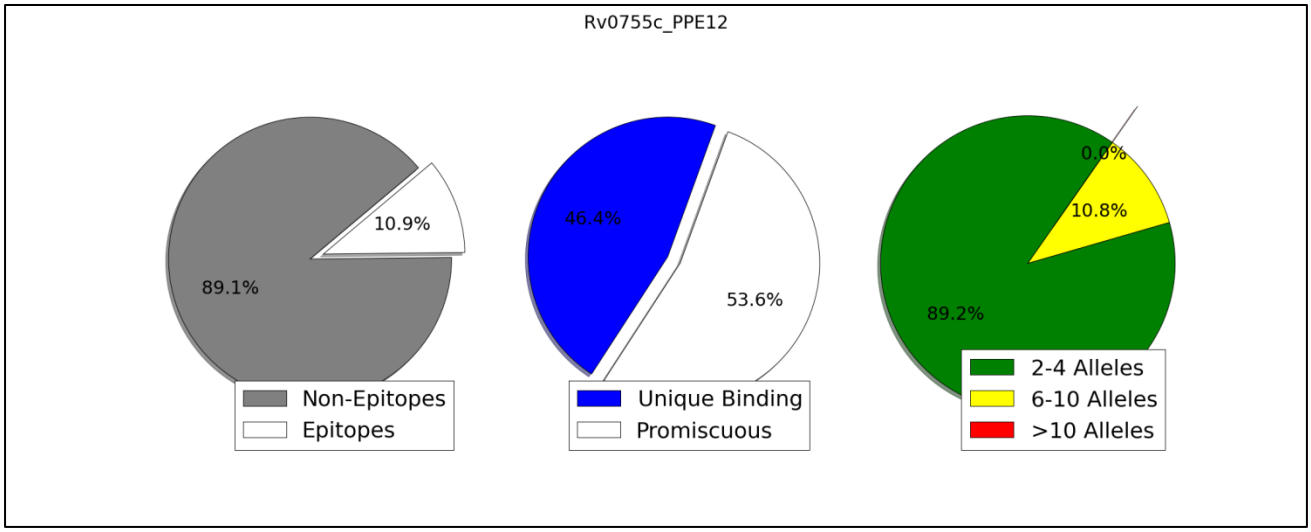
Strong, M. *et al*., 2006. Toward the structural genomics of complexes: crystal structure of a PE/PPE protein complex from *Mycobacterium tuberculosis*. *Proceedings of the National Academy of Sciences of the United States of America*, 103(21), pp.8060–5.

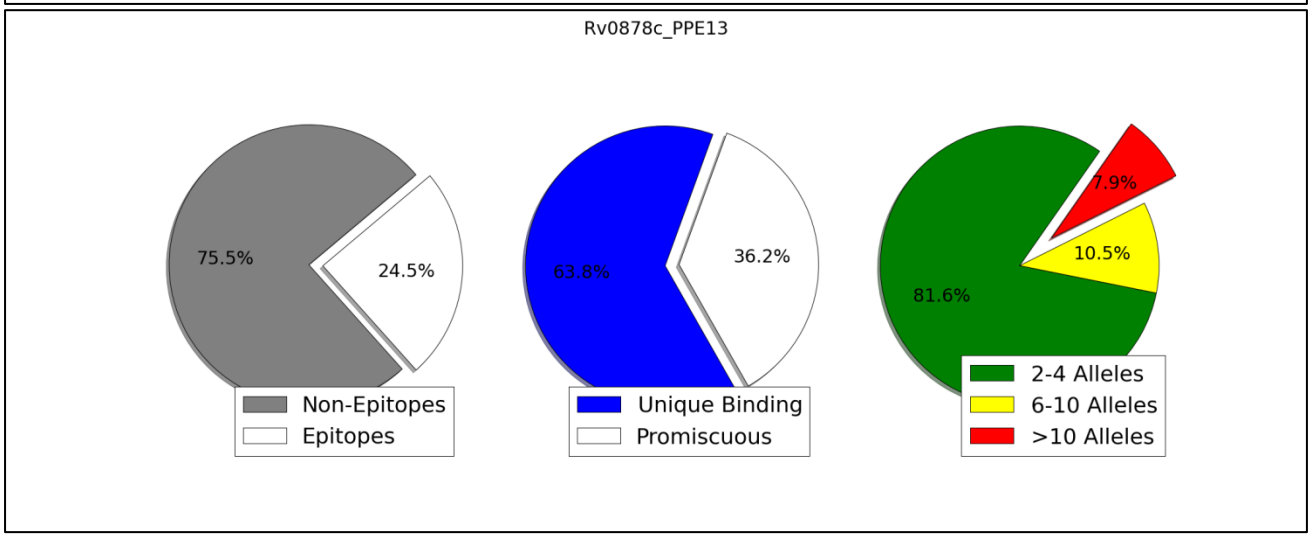
WHO, 2014. Global Tuberculosis Report 2014. Available at: http://reliefweb.int/report/world/global-tuberculosis-report-2014 [Accessed March 29, 2016].

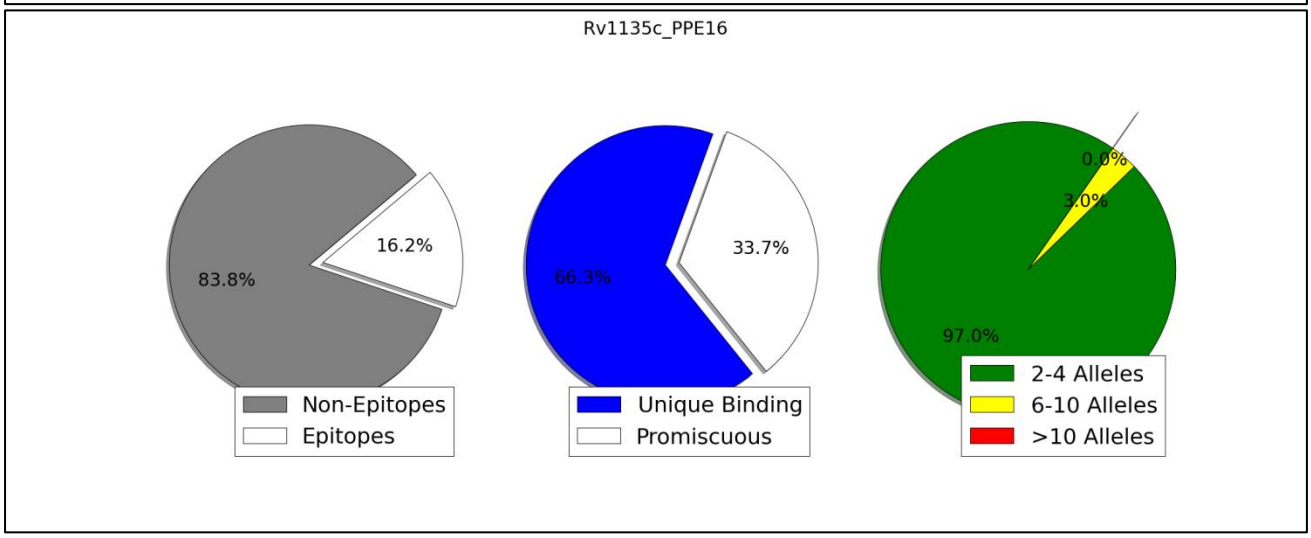
4.6 Appendices

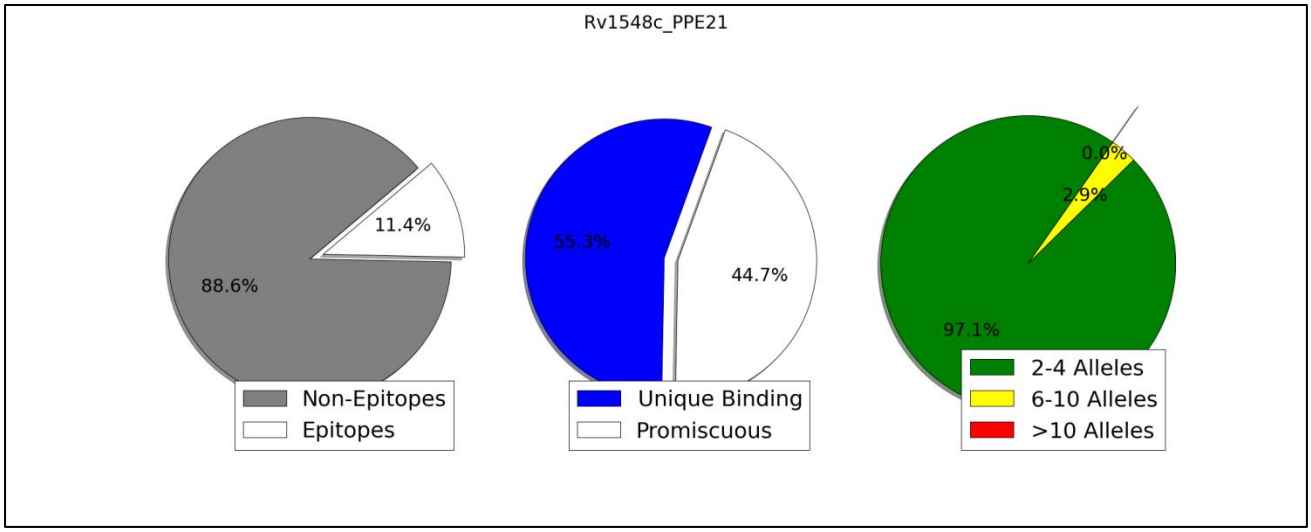
A: Proportion of epitopes versus non-epitopes, unique binders versus promiscuous epitopes, and distribution of promiscuous epitopes.

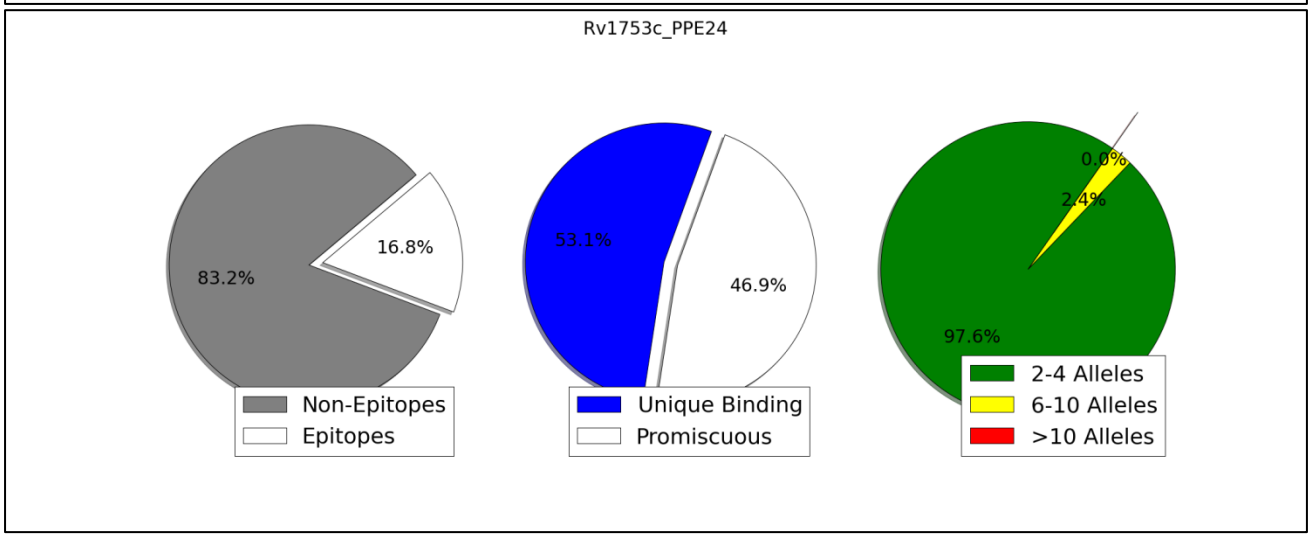


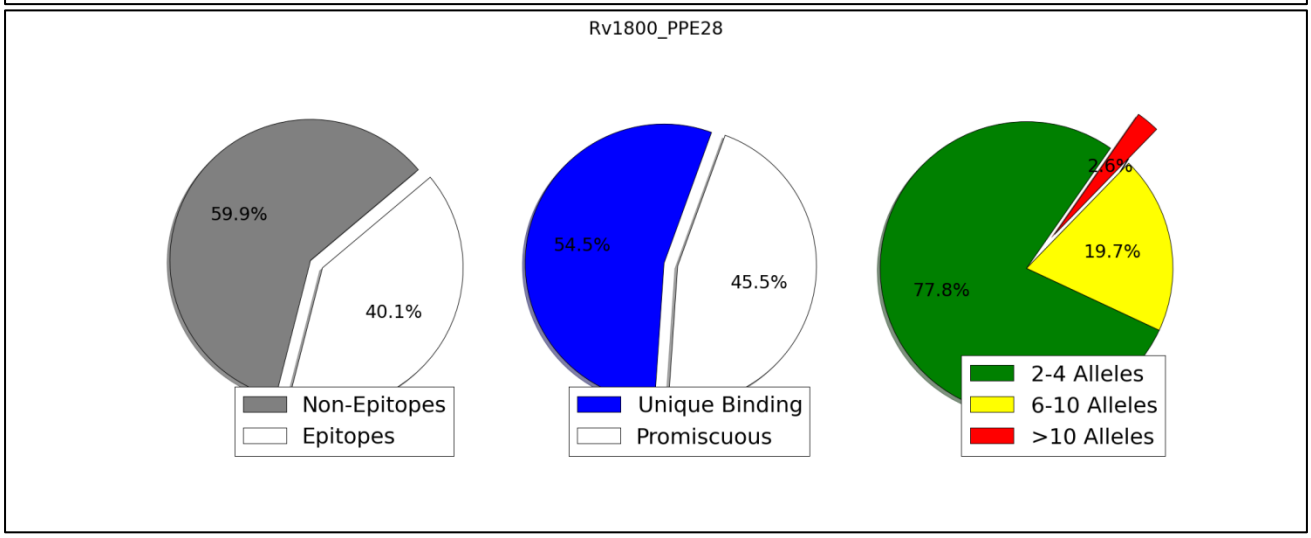


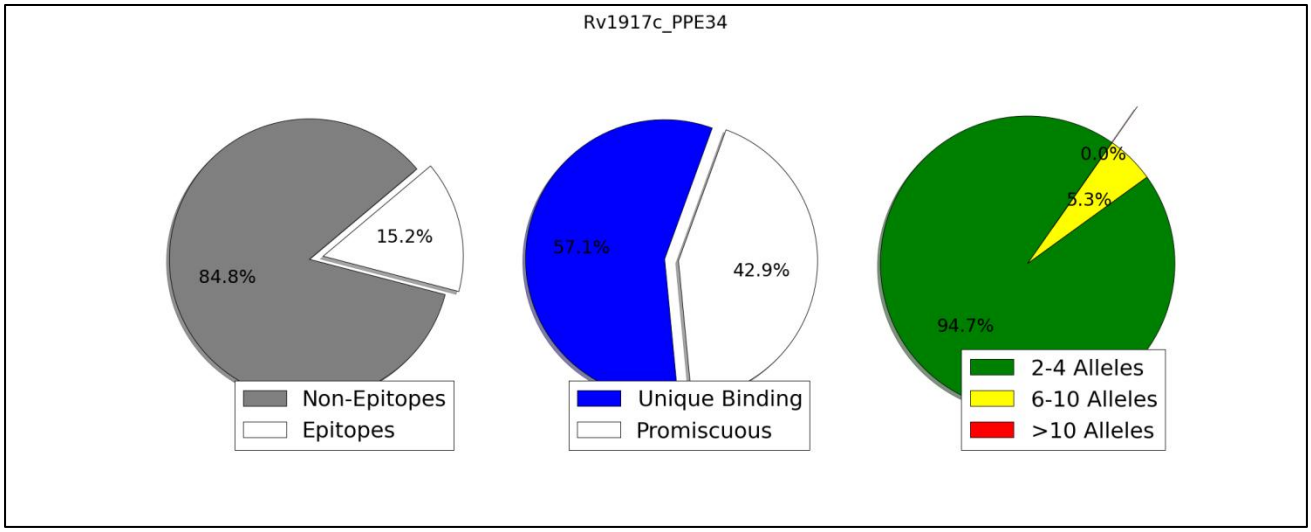


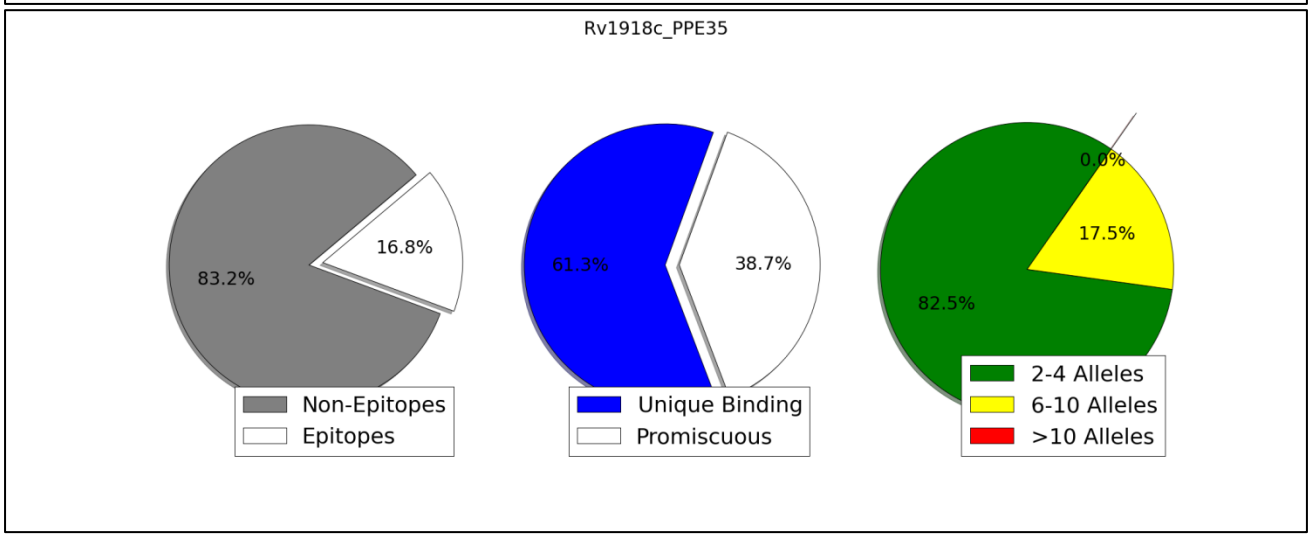


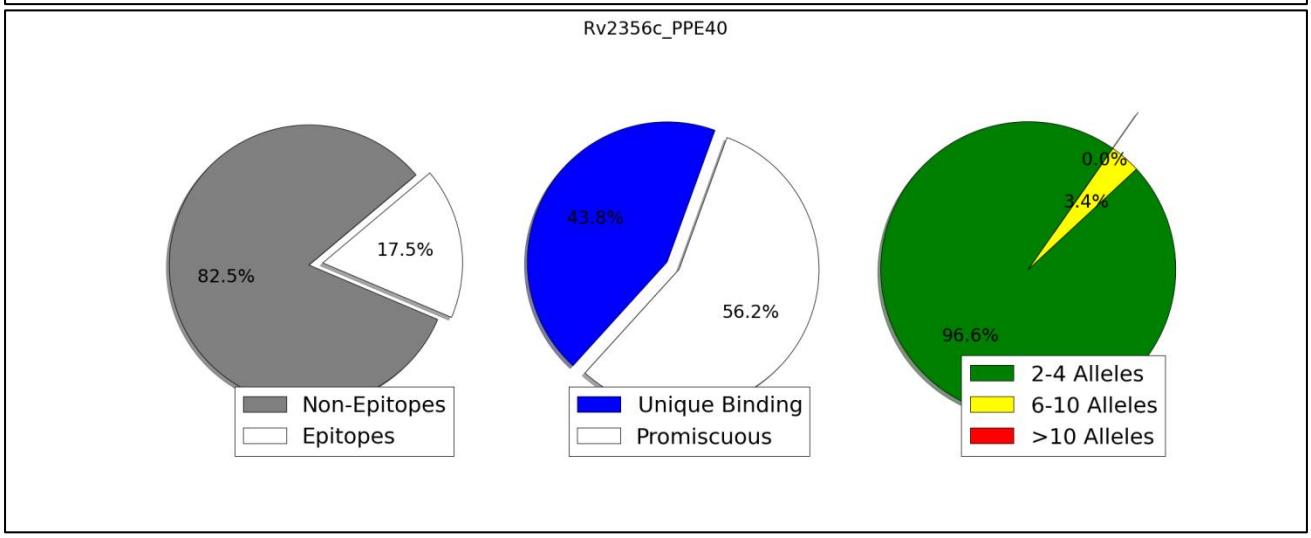


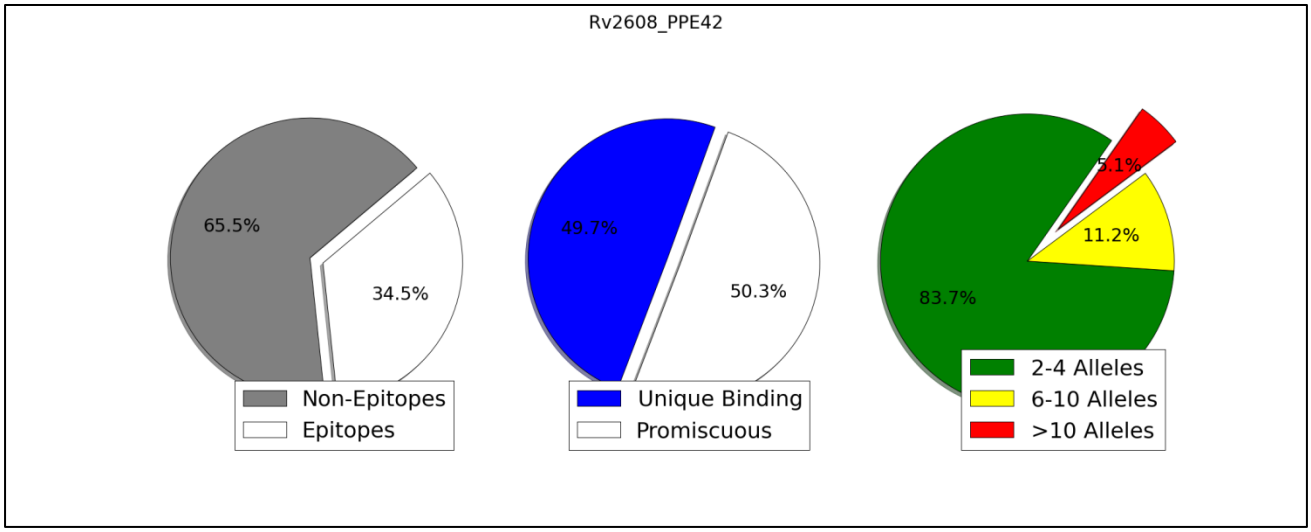


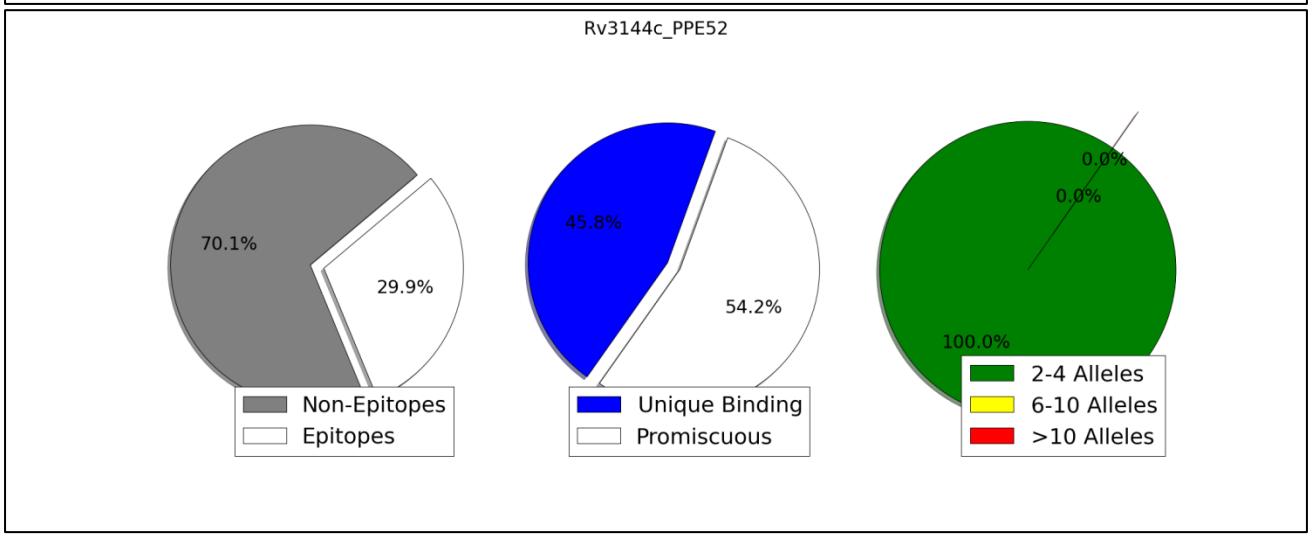


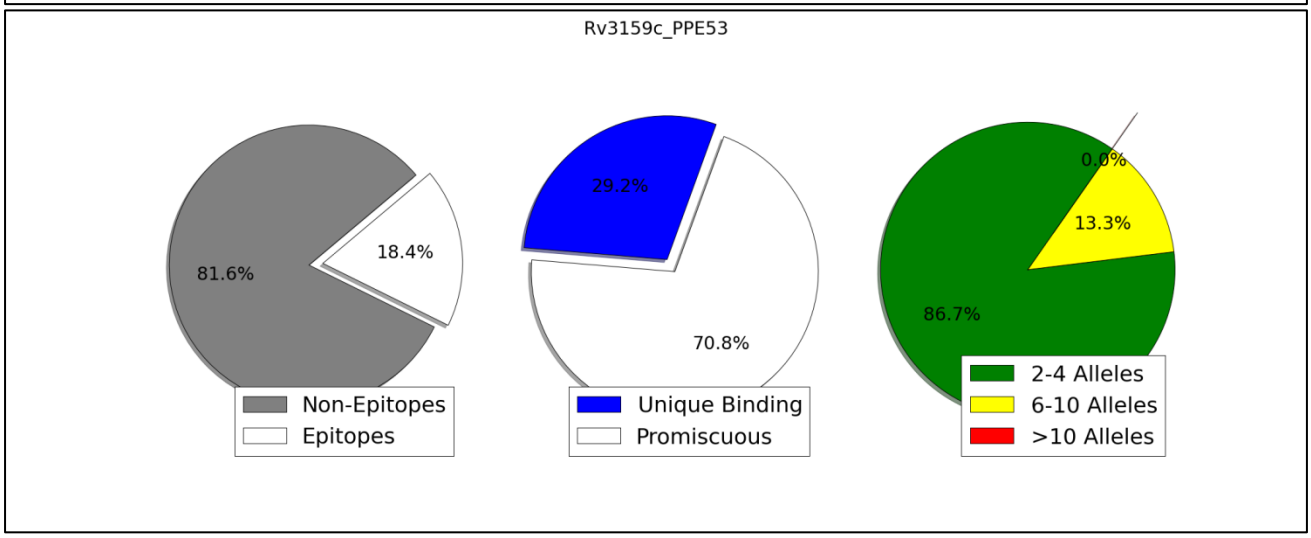


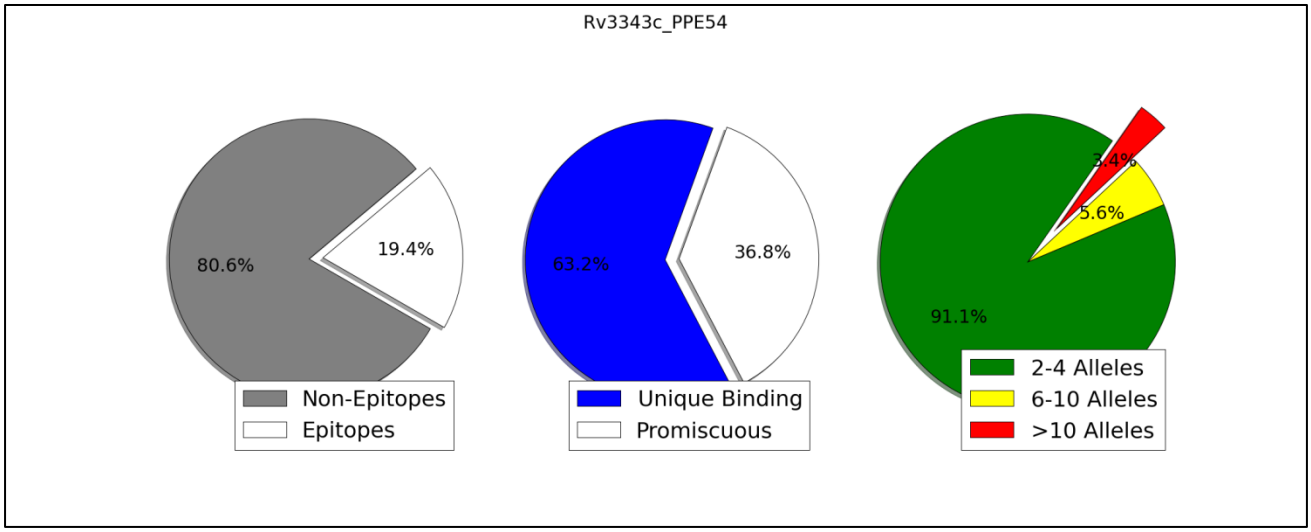


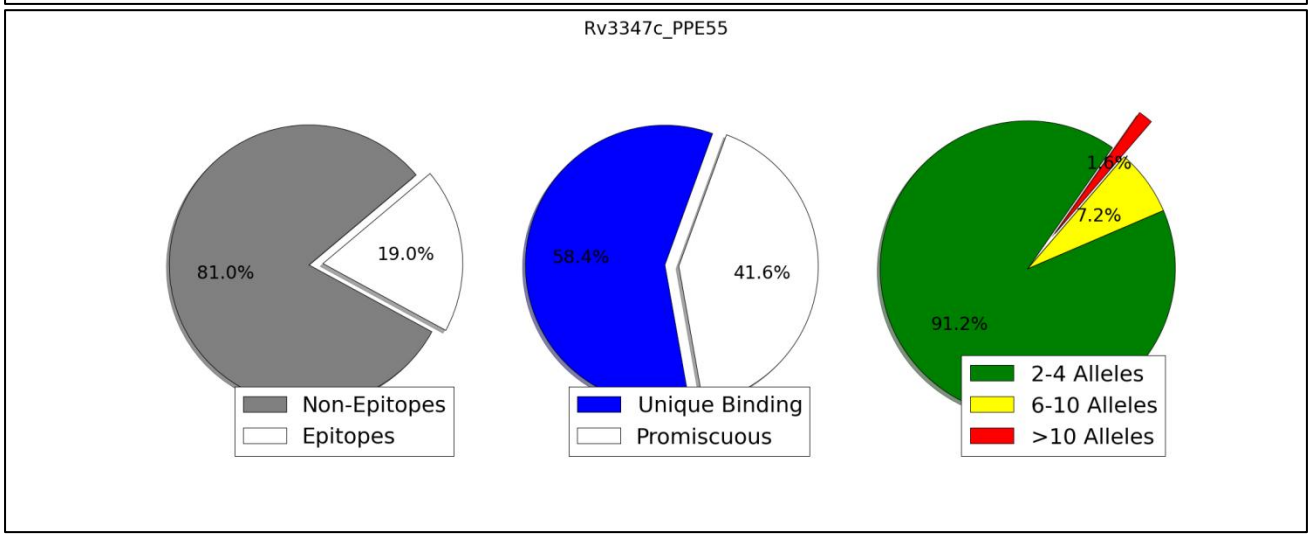


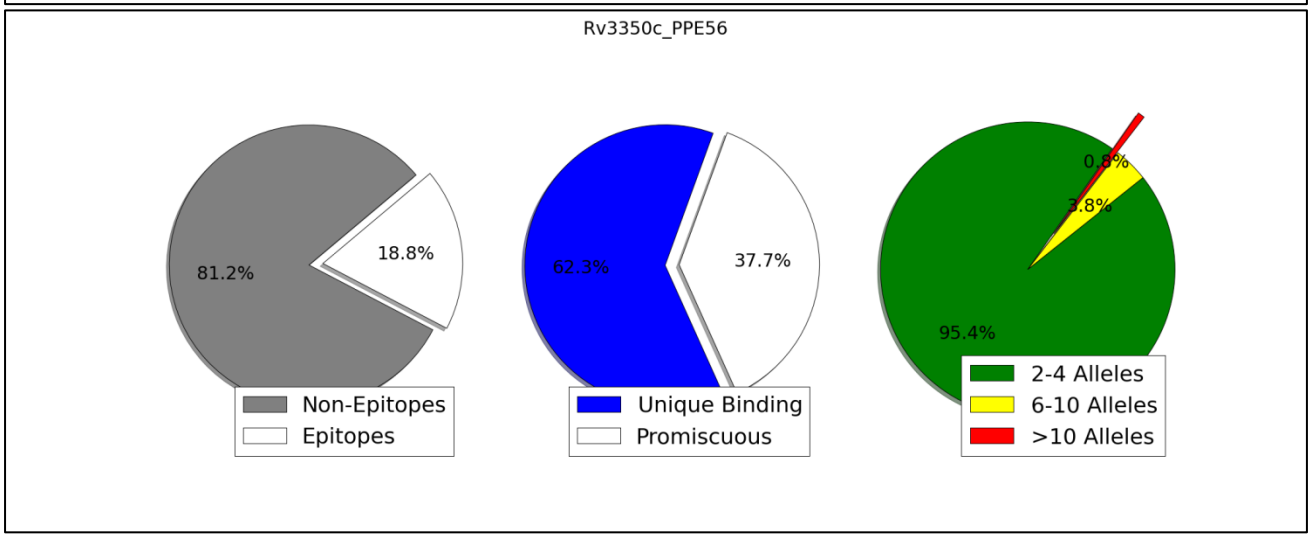


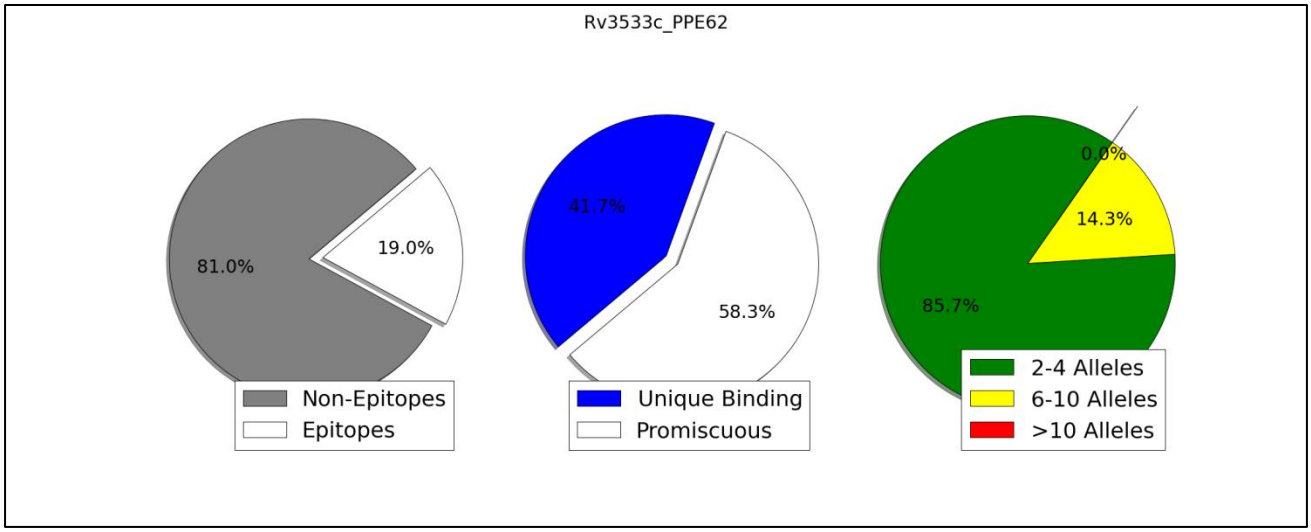


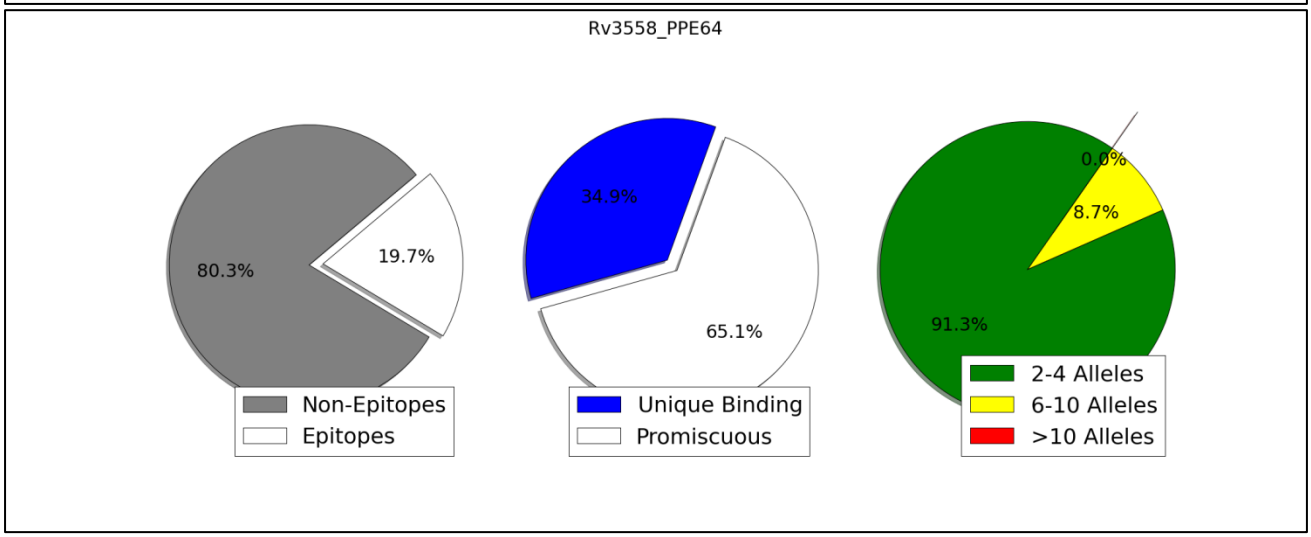








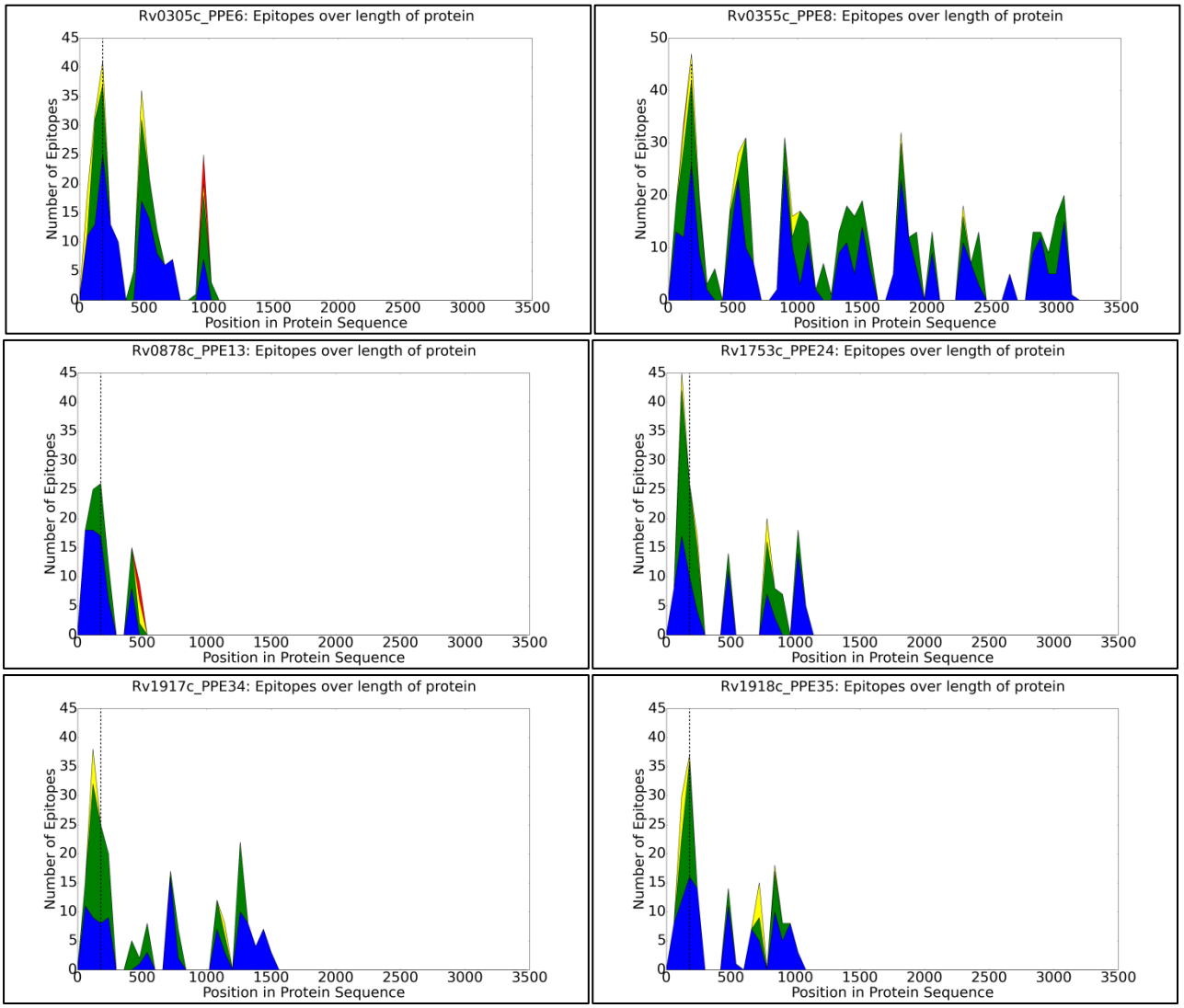


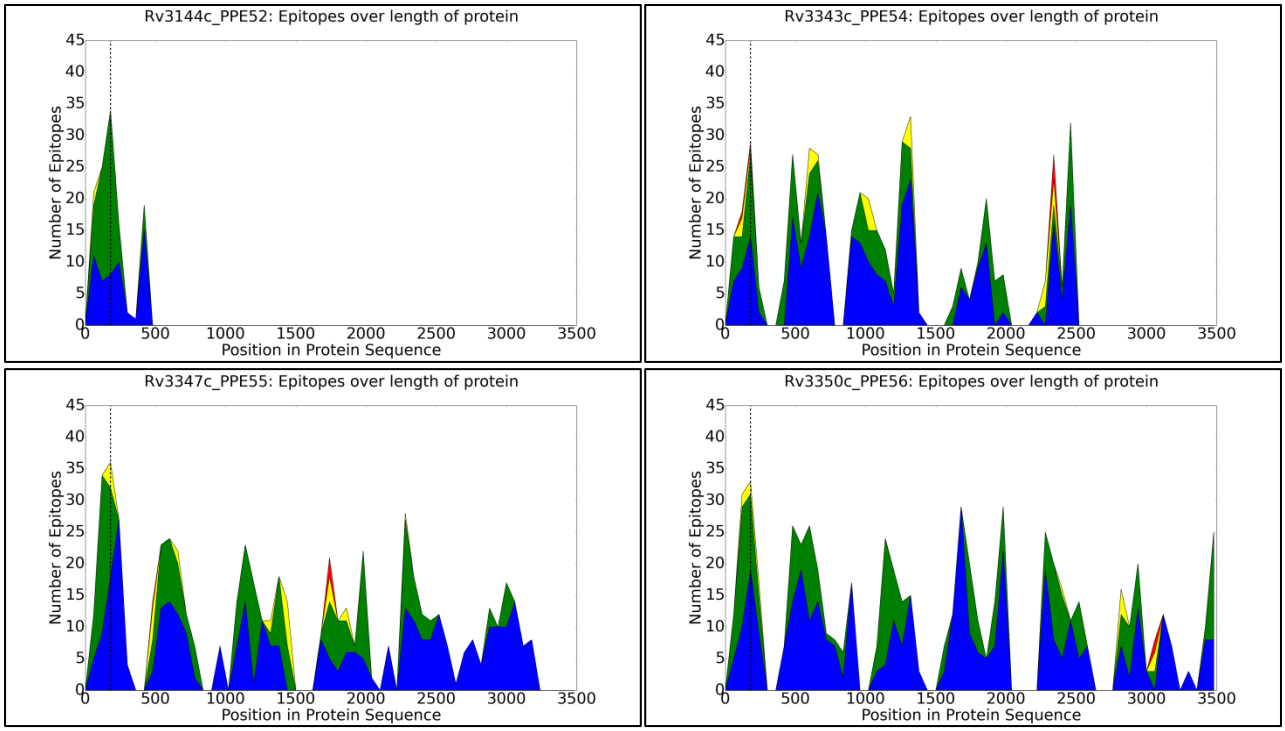


B: Patterns of predicted epitopes showing distribution of promiscuity along length of each protein.

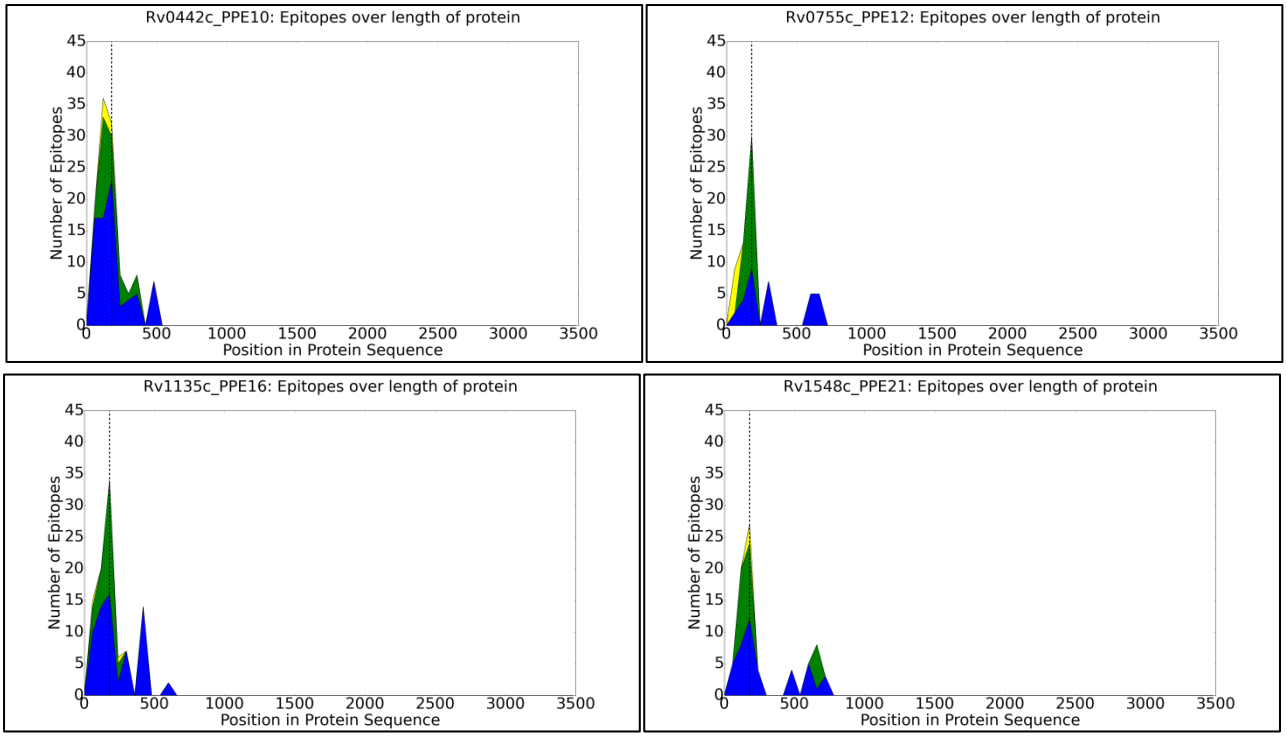
The number of predicted epitopes along the length of each protein is shown, indicating the pattern of fluctuation (fluctuating, decreasing or increasing/stable). Colours in the graph represent the number of HLA alleles each epitope is predicted to bind to. (Blue: 1 allele, Green: 2-4 alleles, Yellow: 5-10 alleles, Red: >10 alleles).

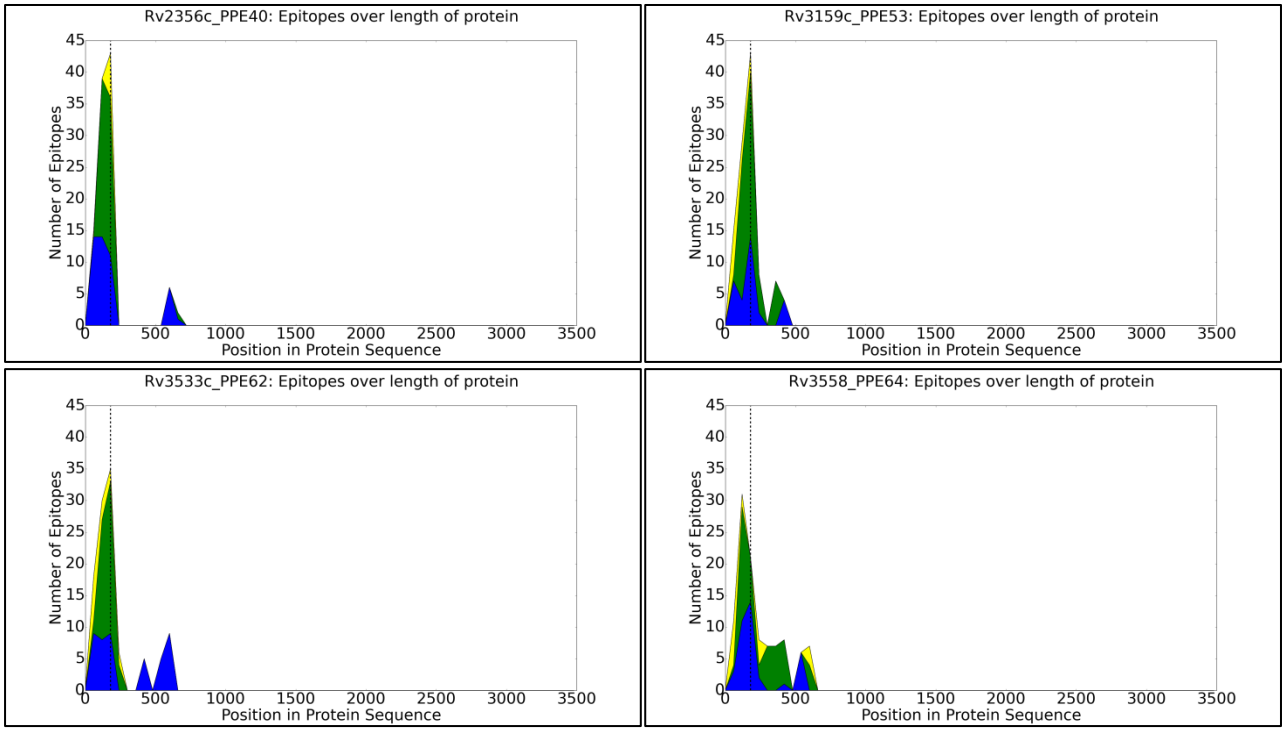
Fluctuating:



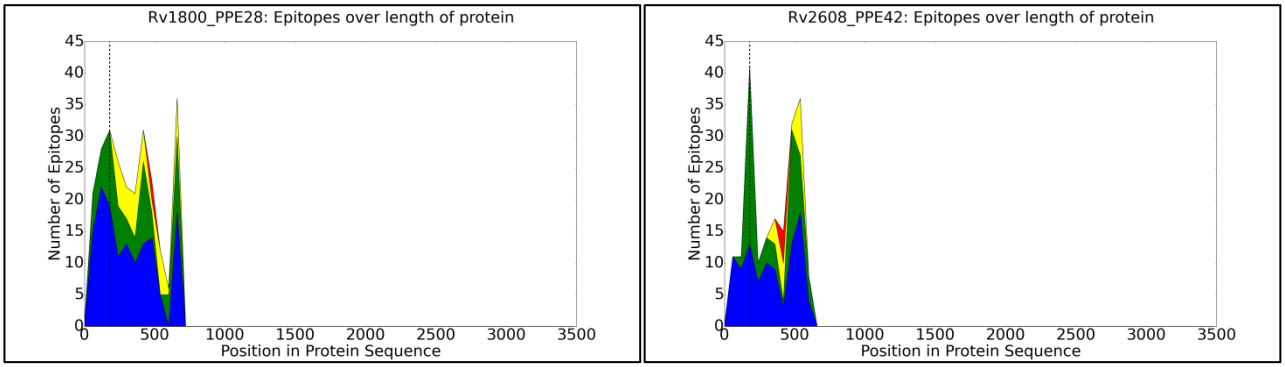


Decreasing:



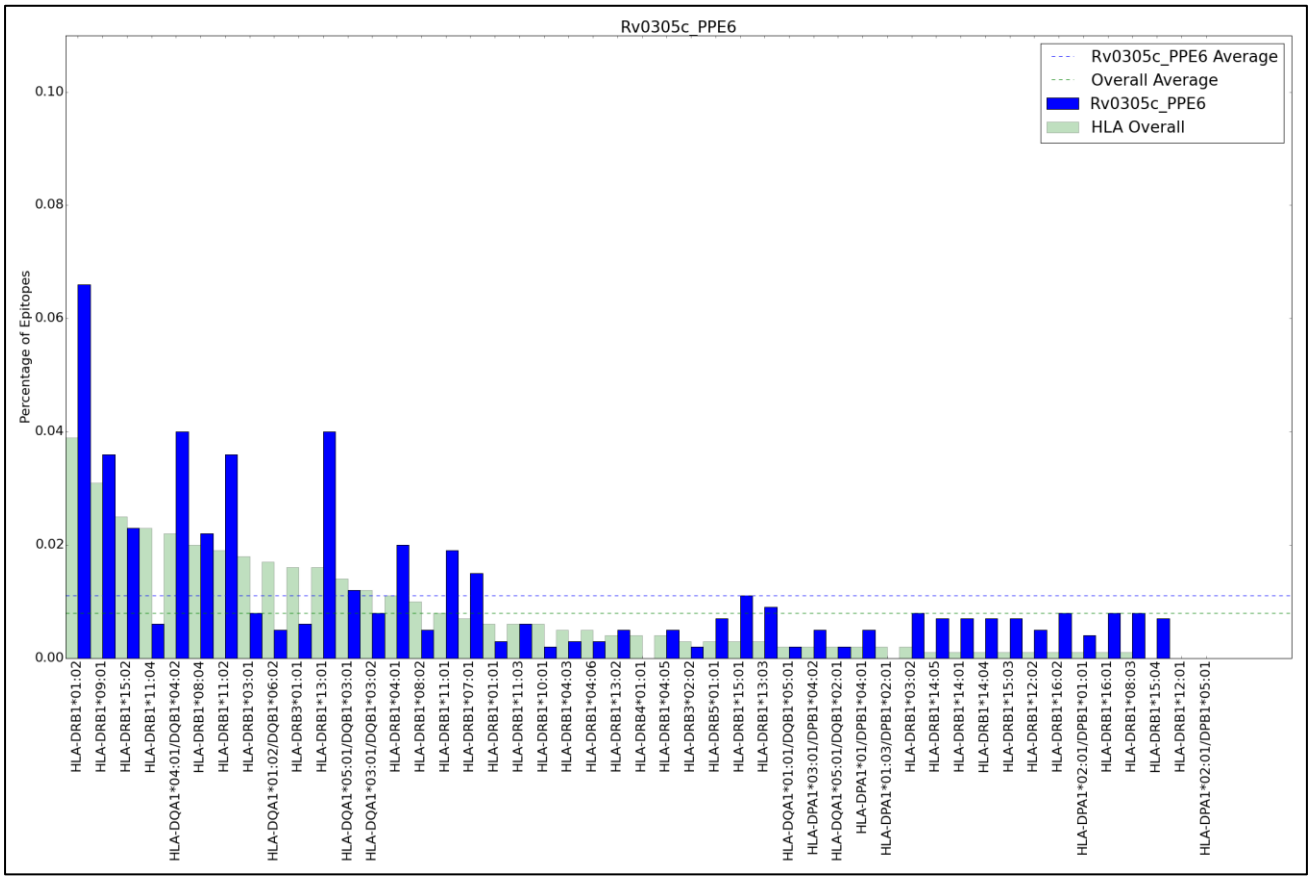


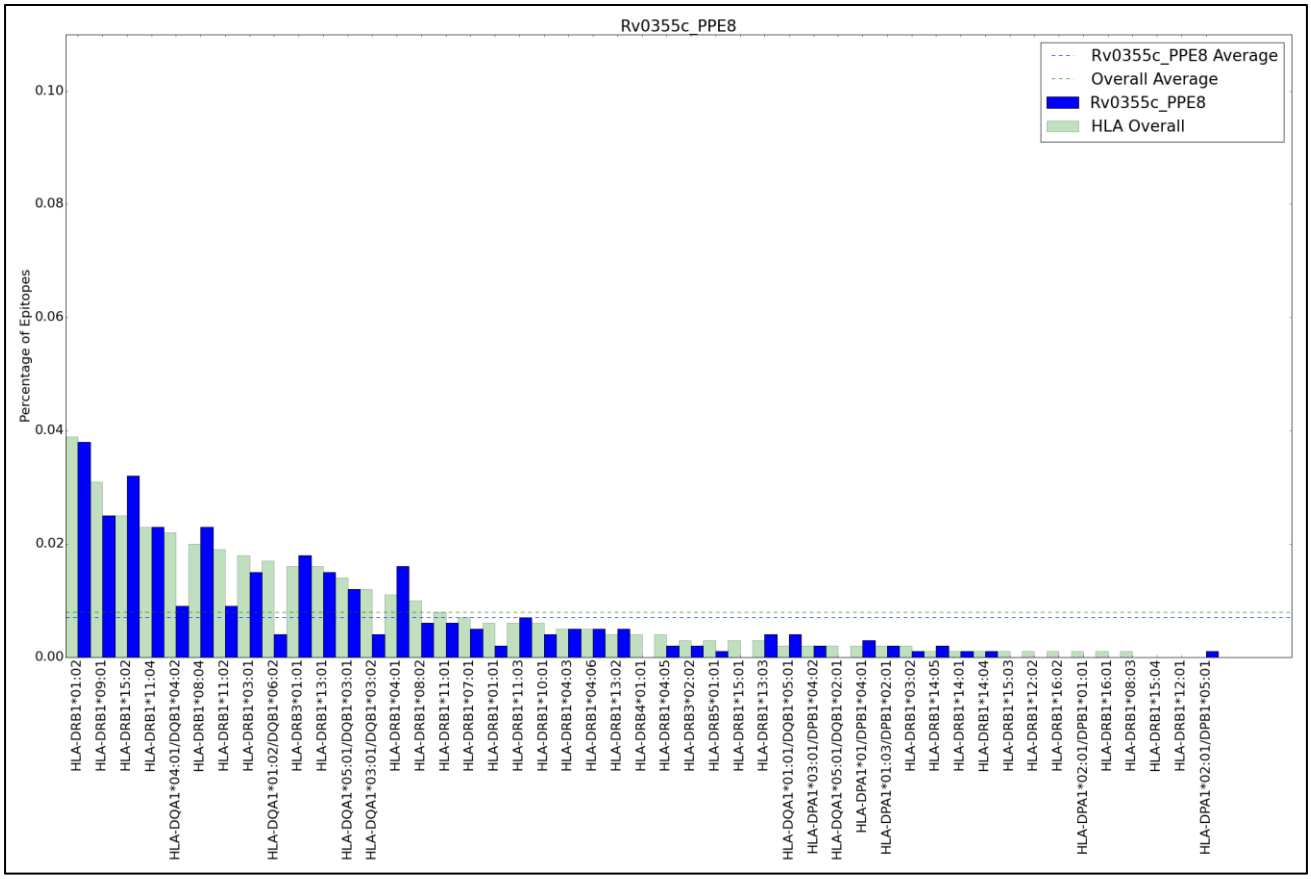
Increasing/Stable:

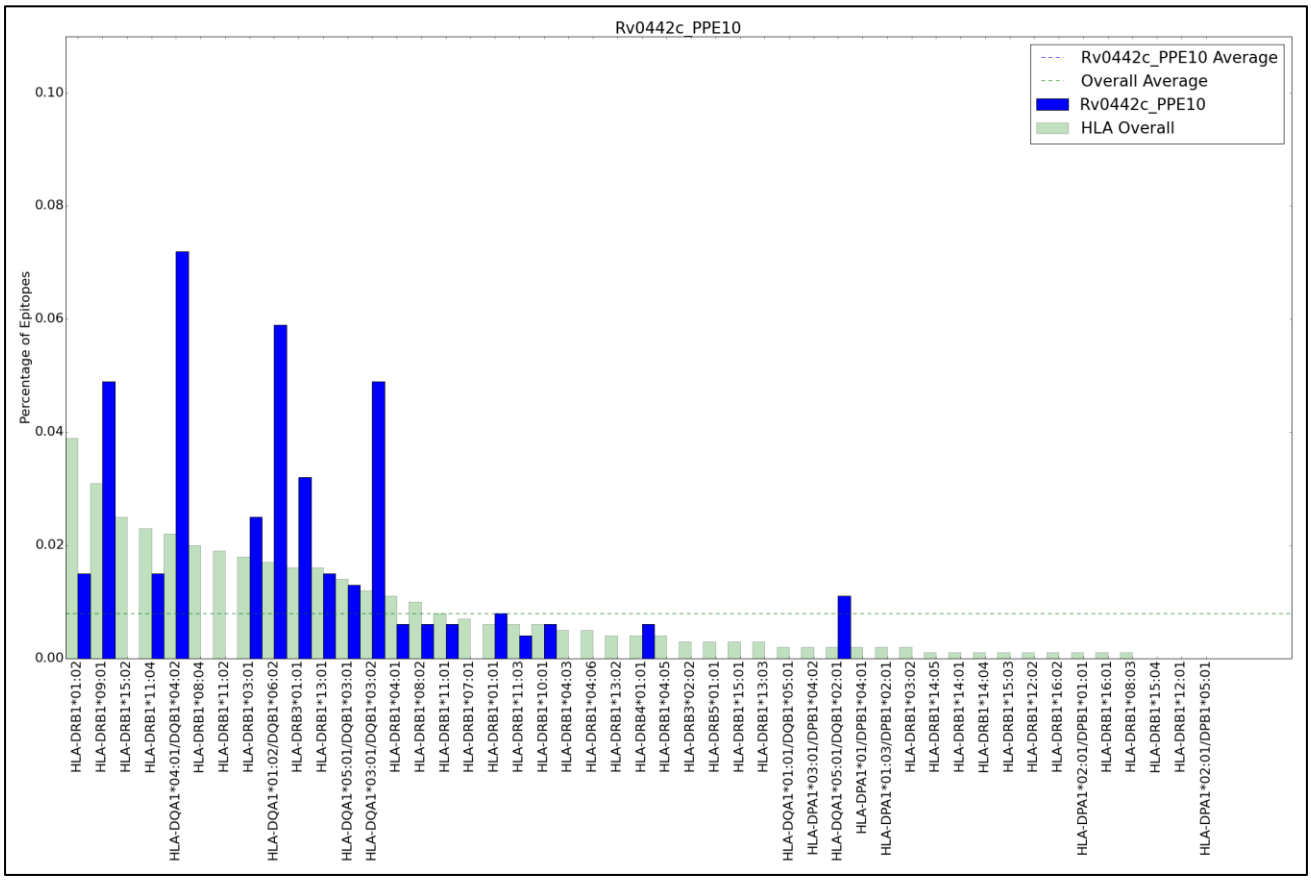


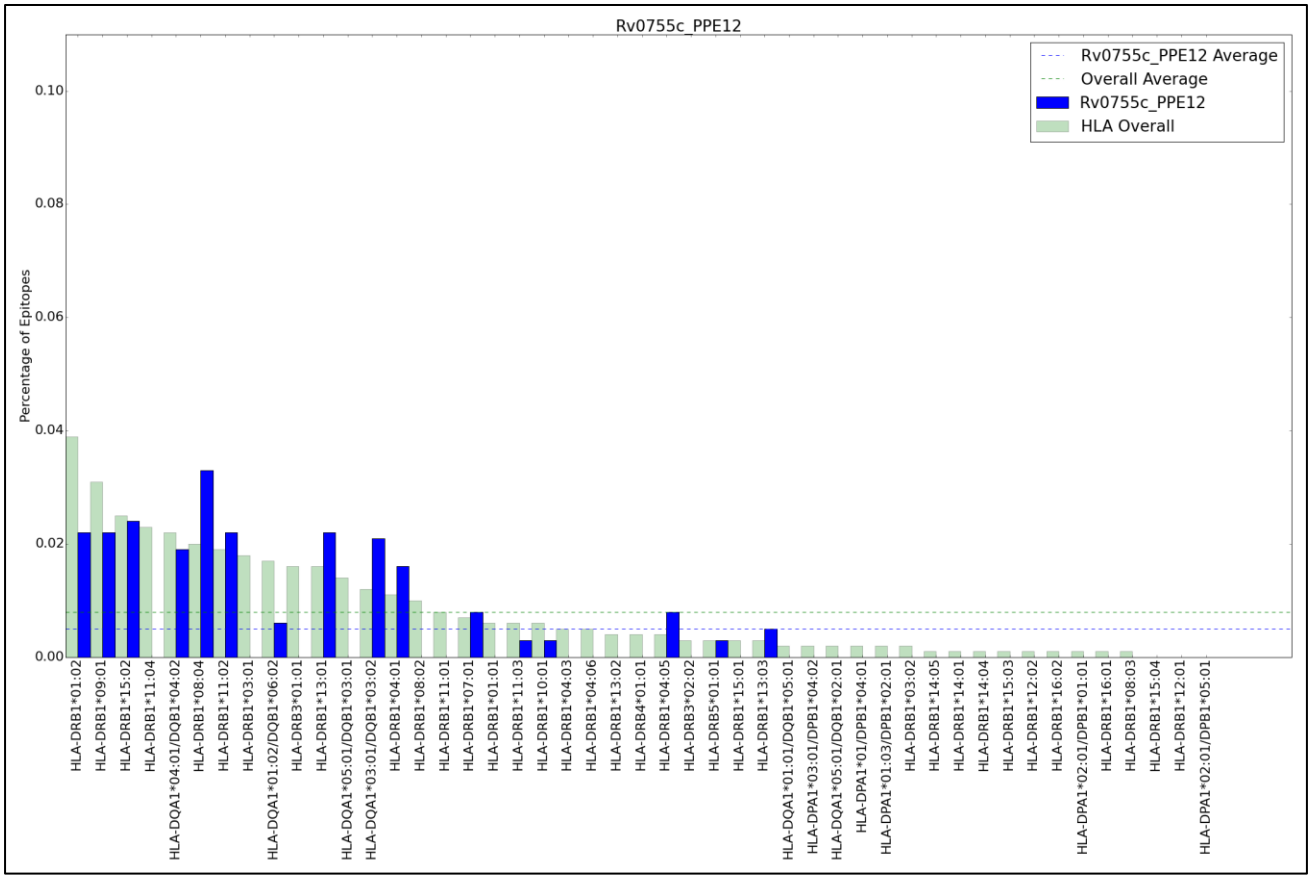
C: Binding ability of HLA alleles per PPE_MPTR protein

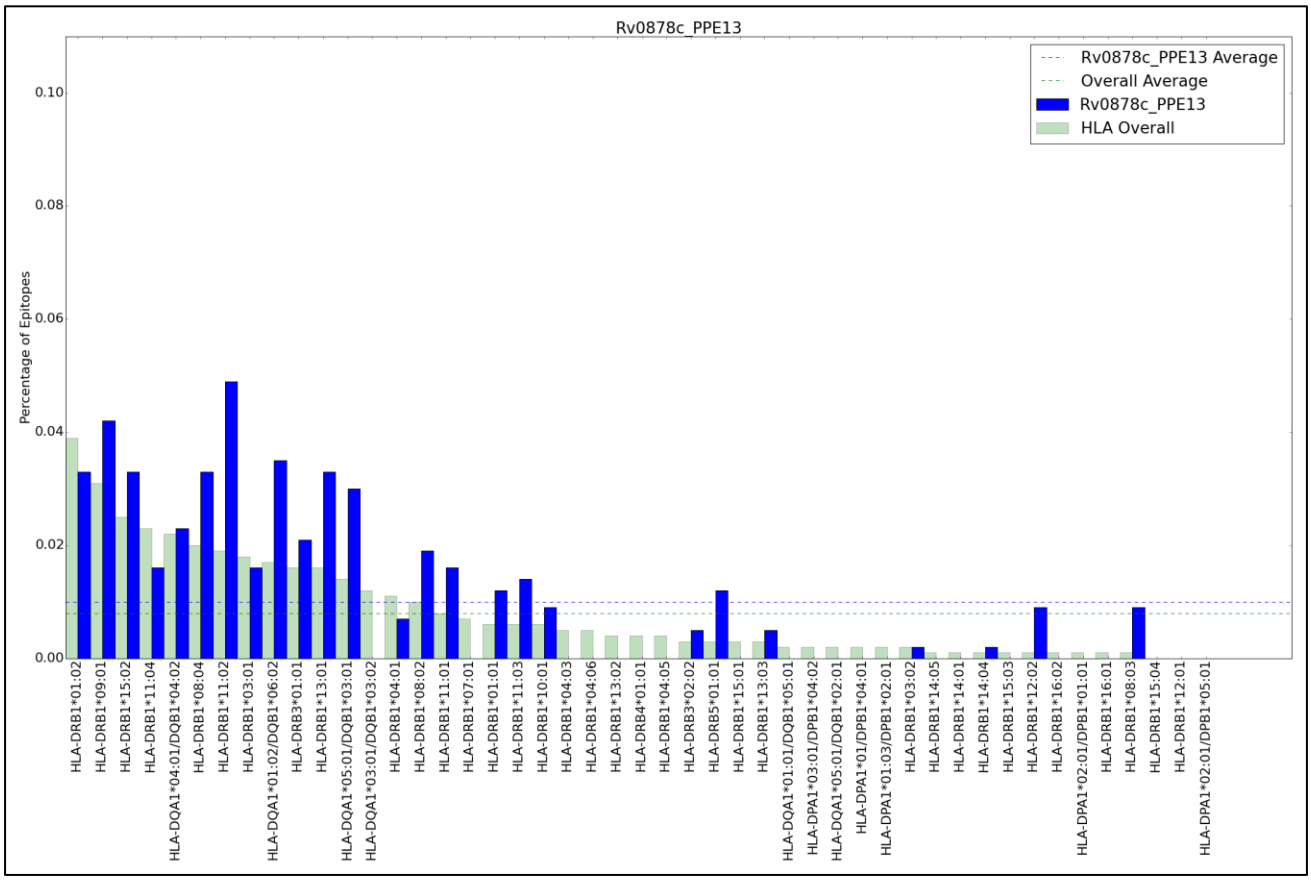
The percentage of peptides each of the HLA alleles are able to bind to was investigated for all of the PPE_MPTR proteins together (shown in green), and for each of the PPE_MPTR proteins individually (show in blue). Dotted lines represent the average percentage of peptides HLA alleles were able to bind to. The distribution across the HLA alleles differs significantly when comparing the different PPE_MPTR proteins, indicating that vaccine candidates derived from different PPE_MPTR proteins will have substantially different population coverage (Chapter 6).

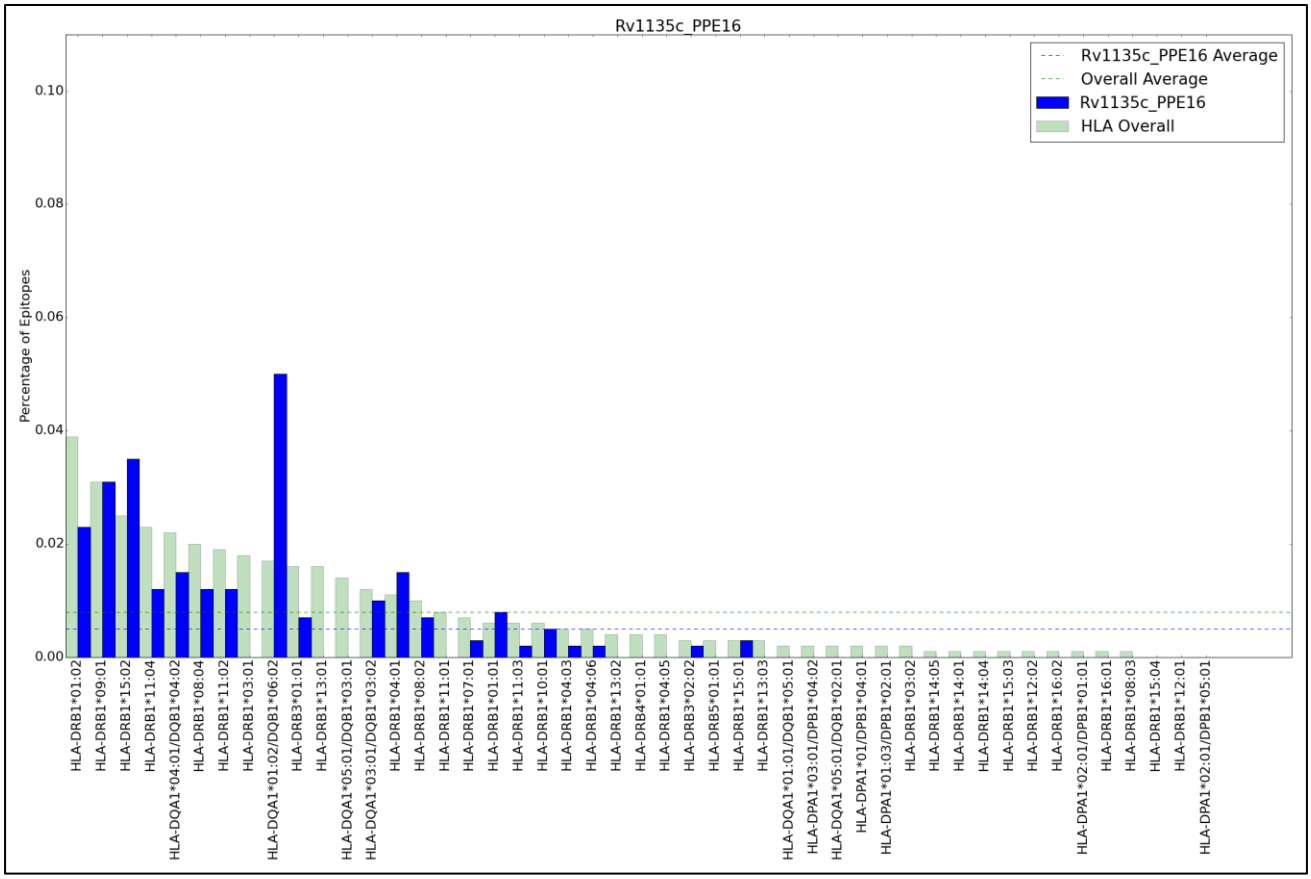


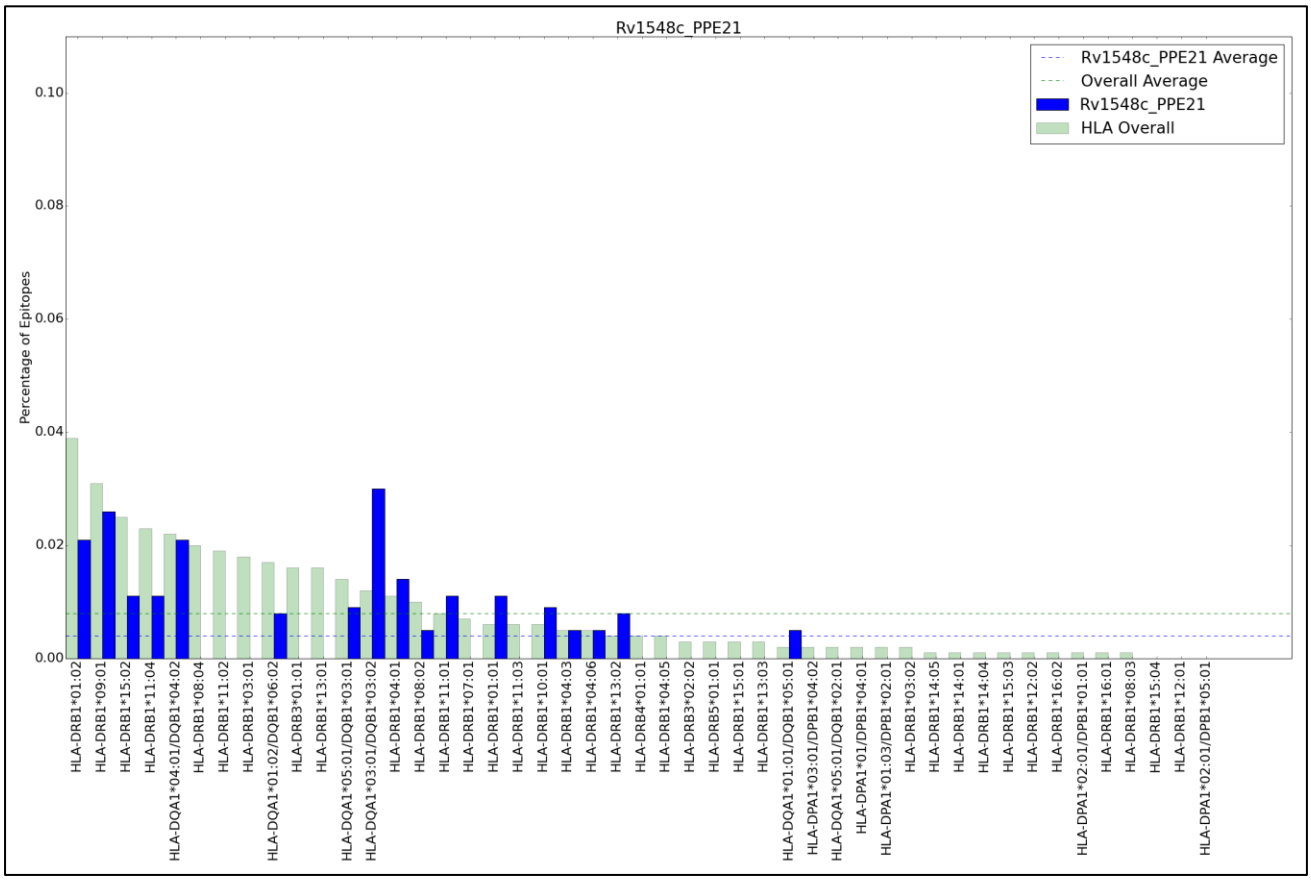


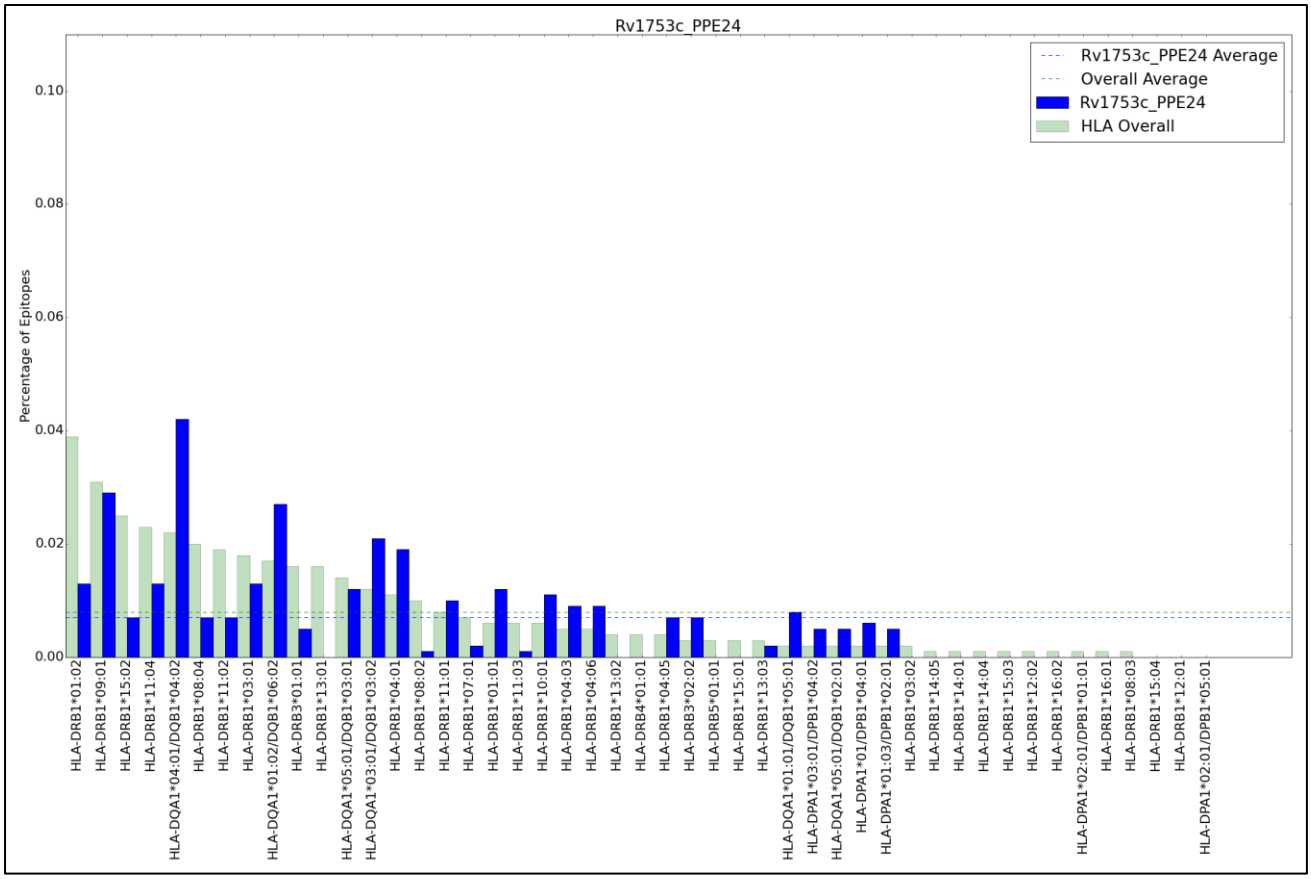


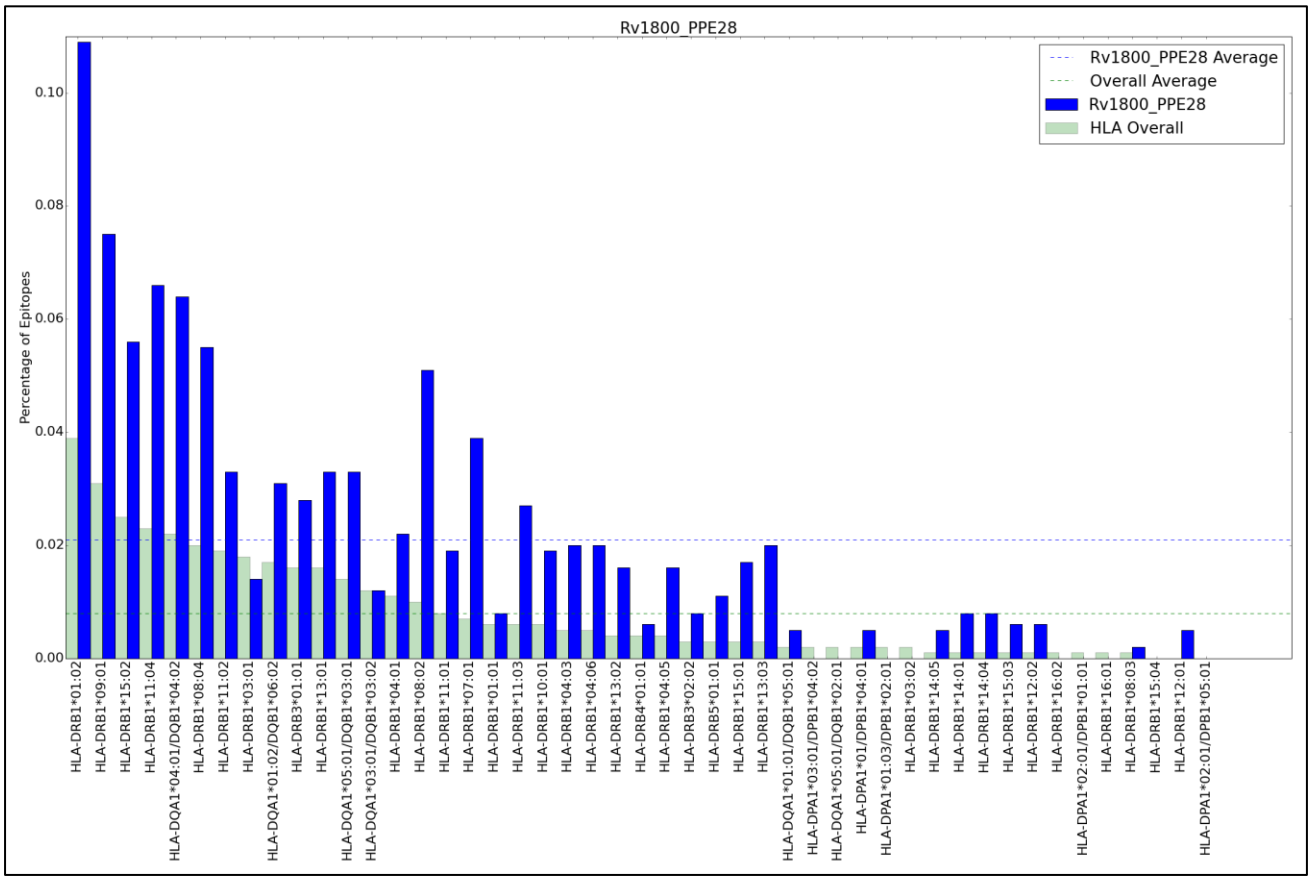


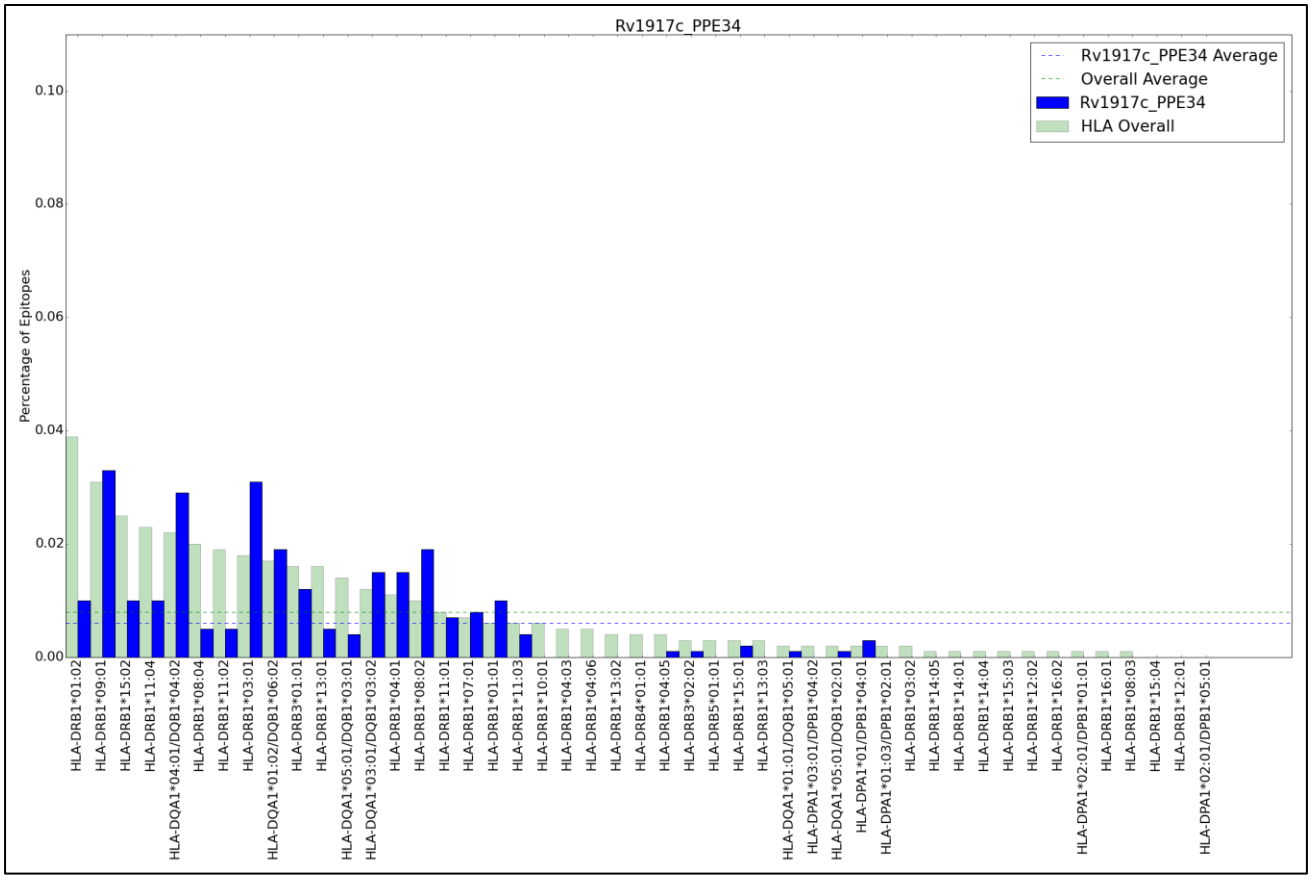


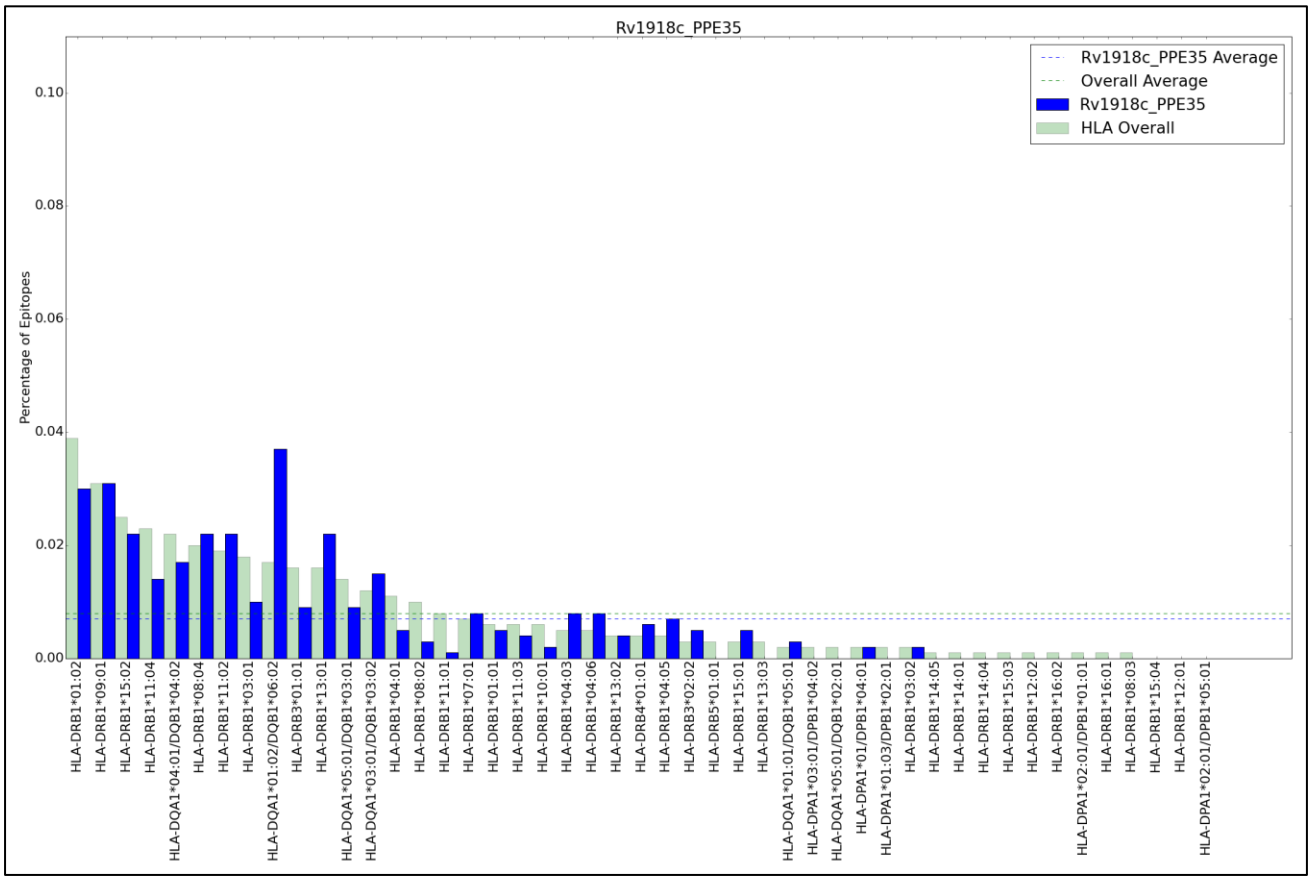


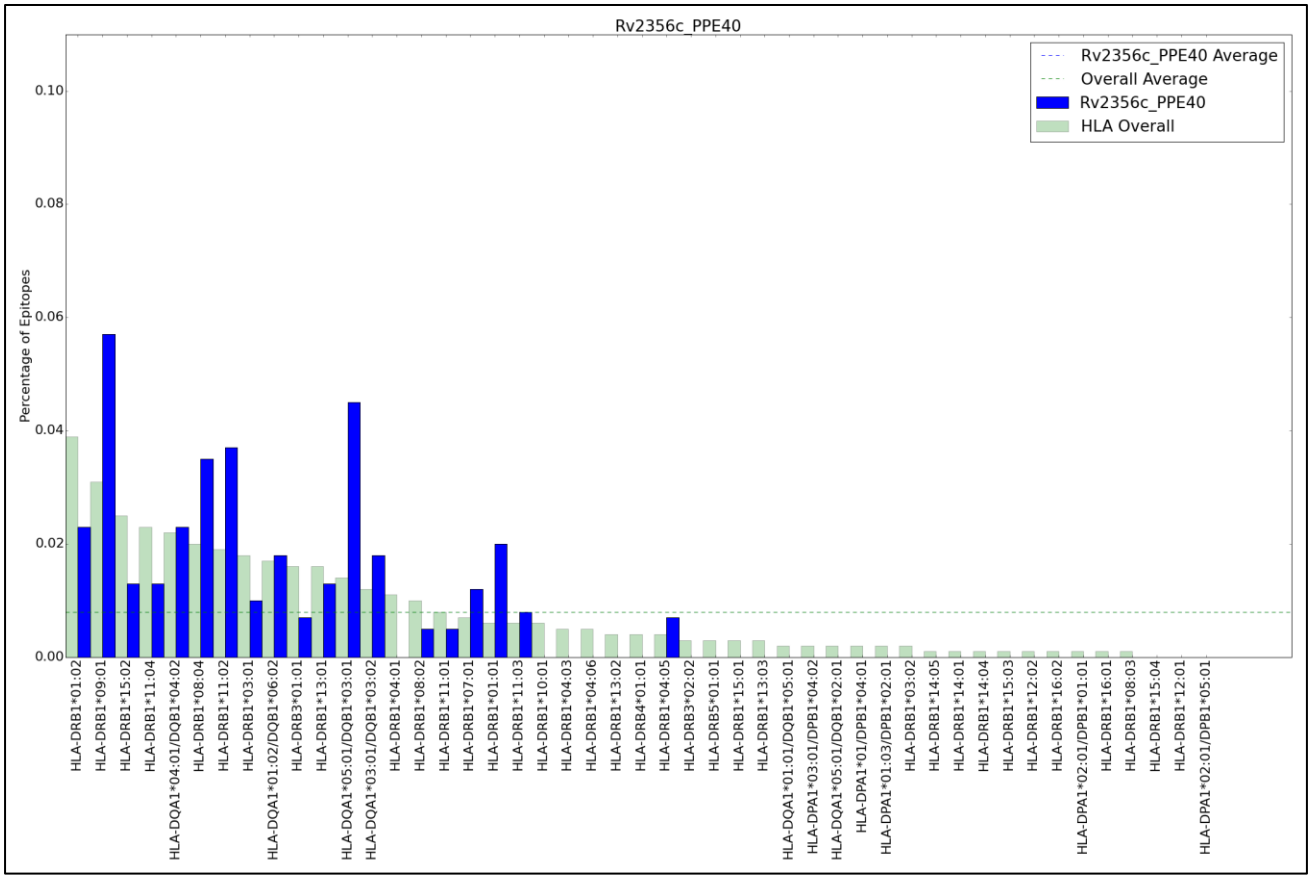


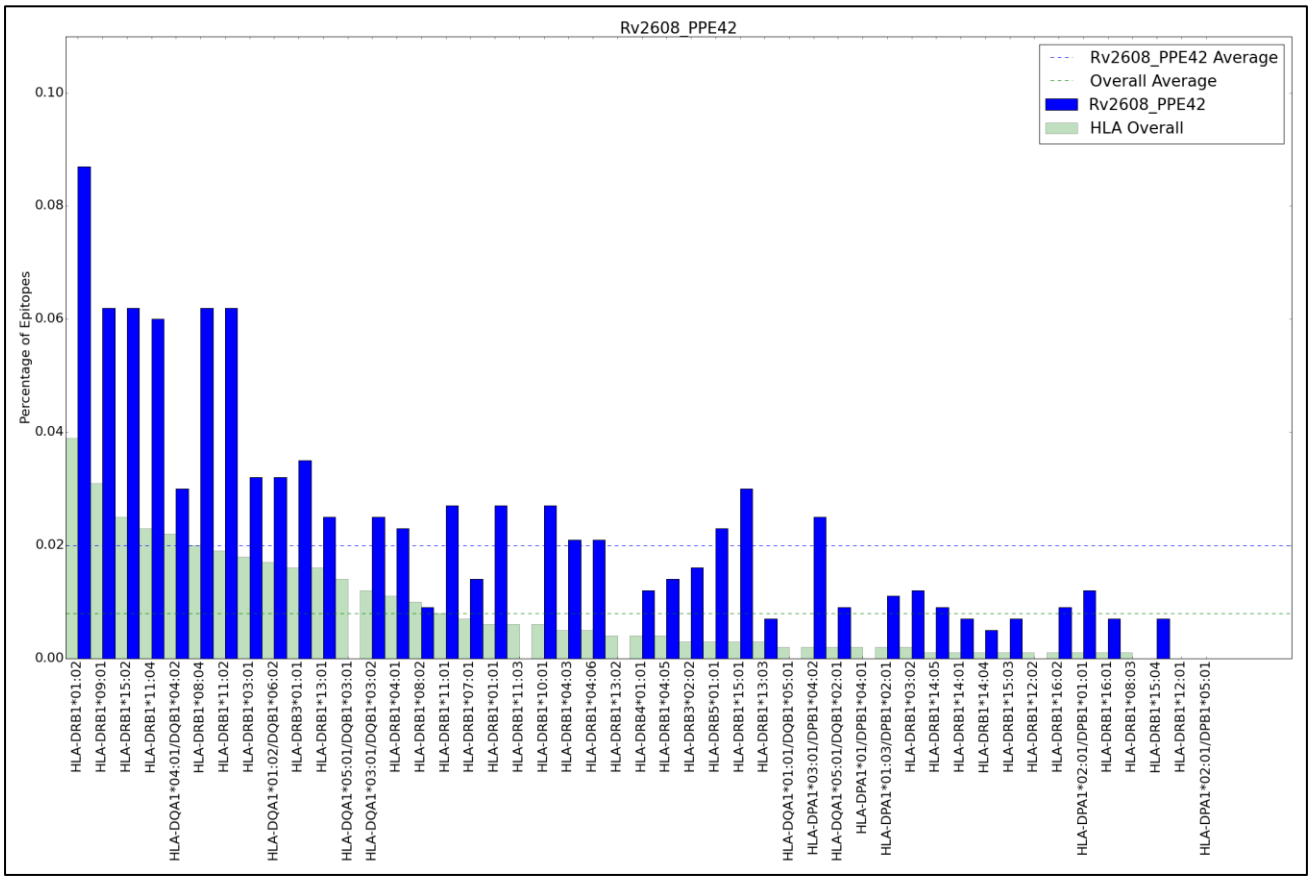


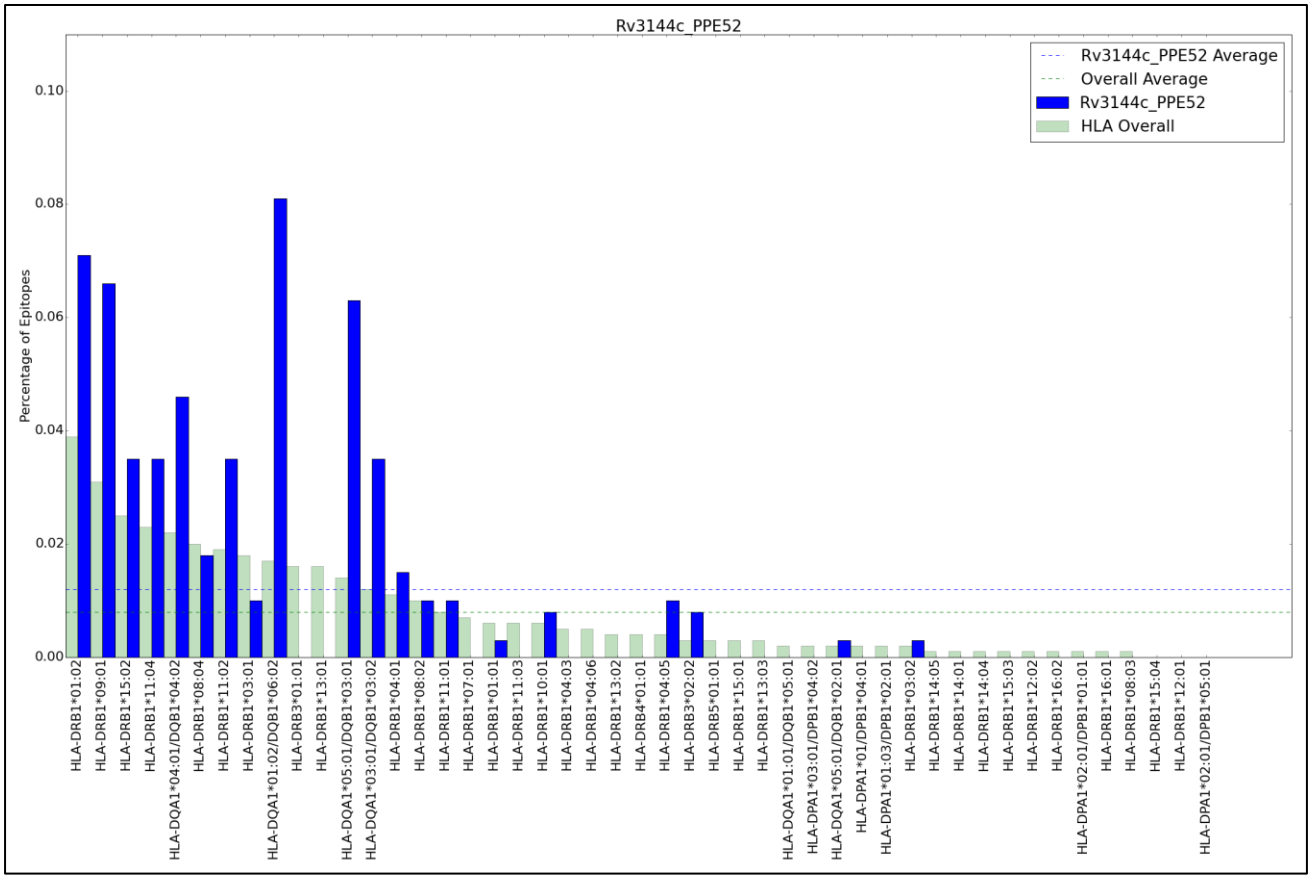


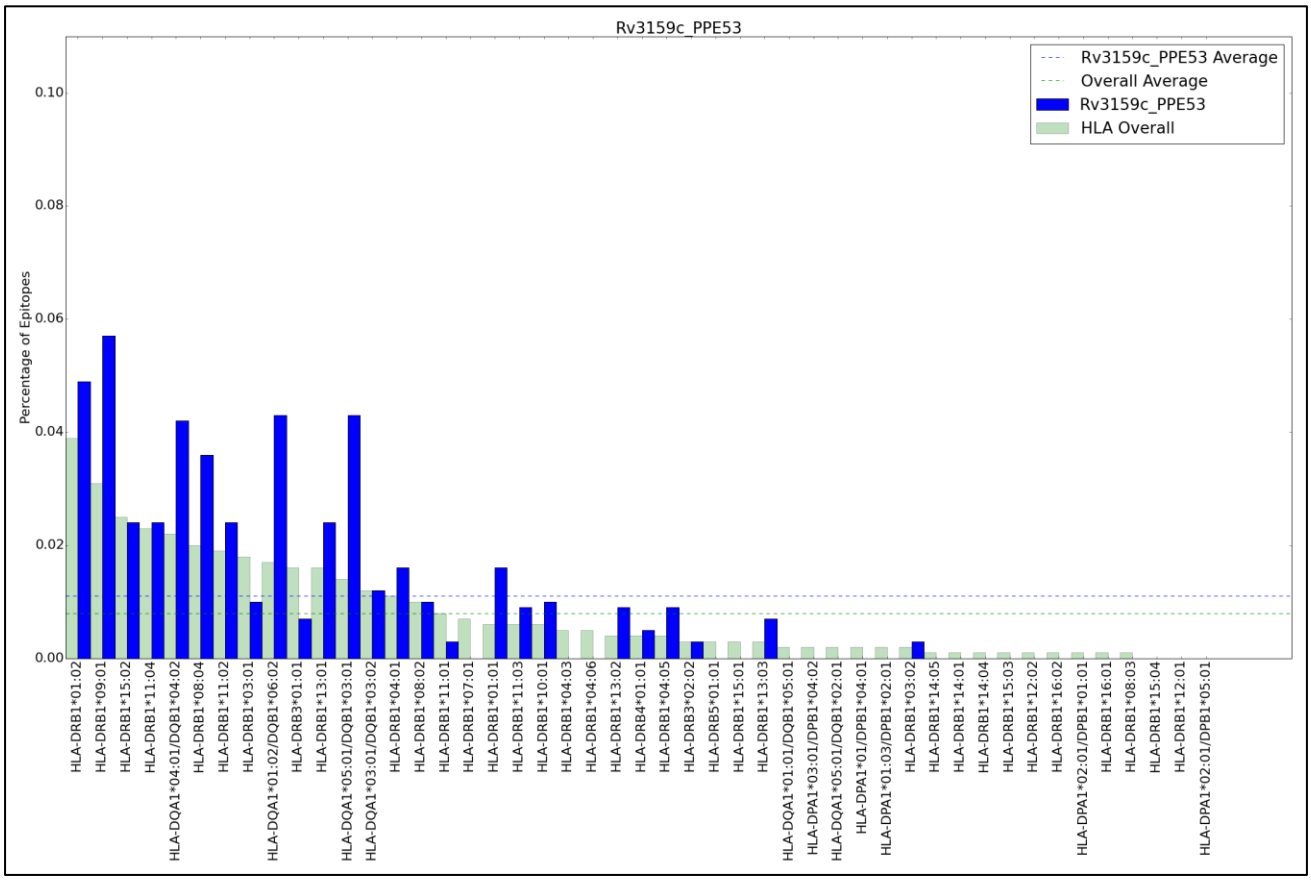


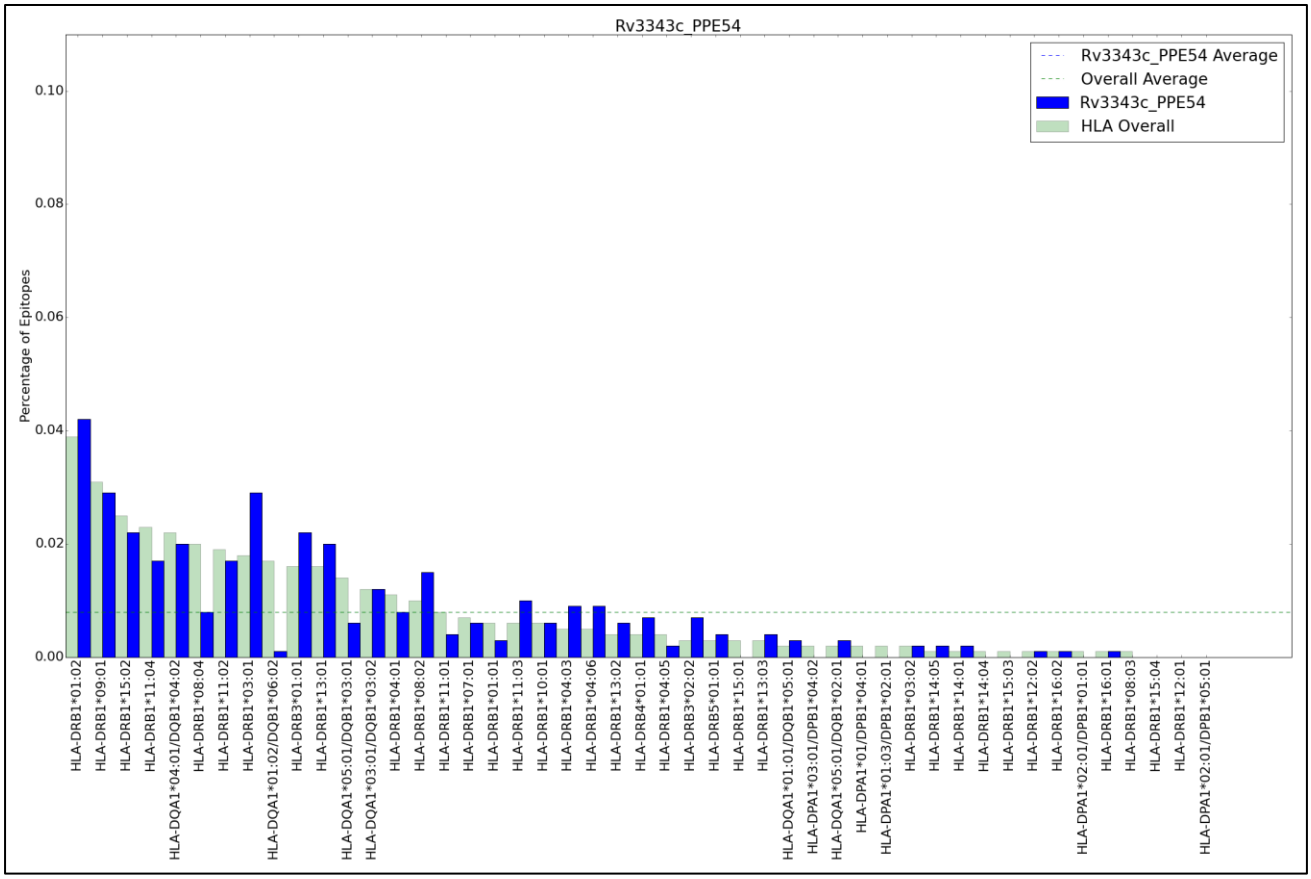


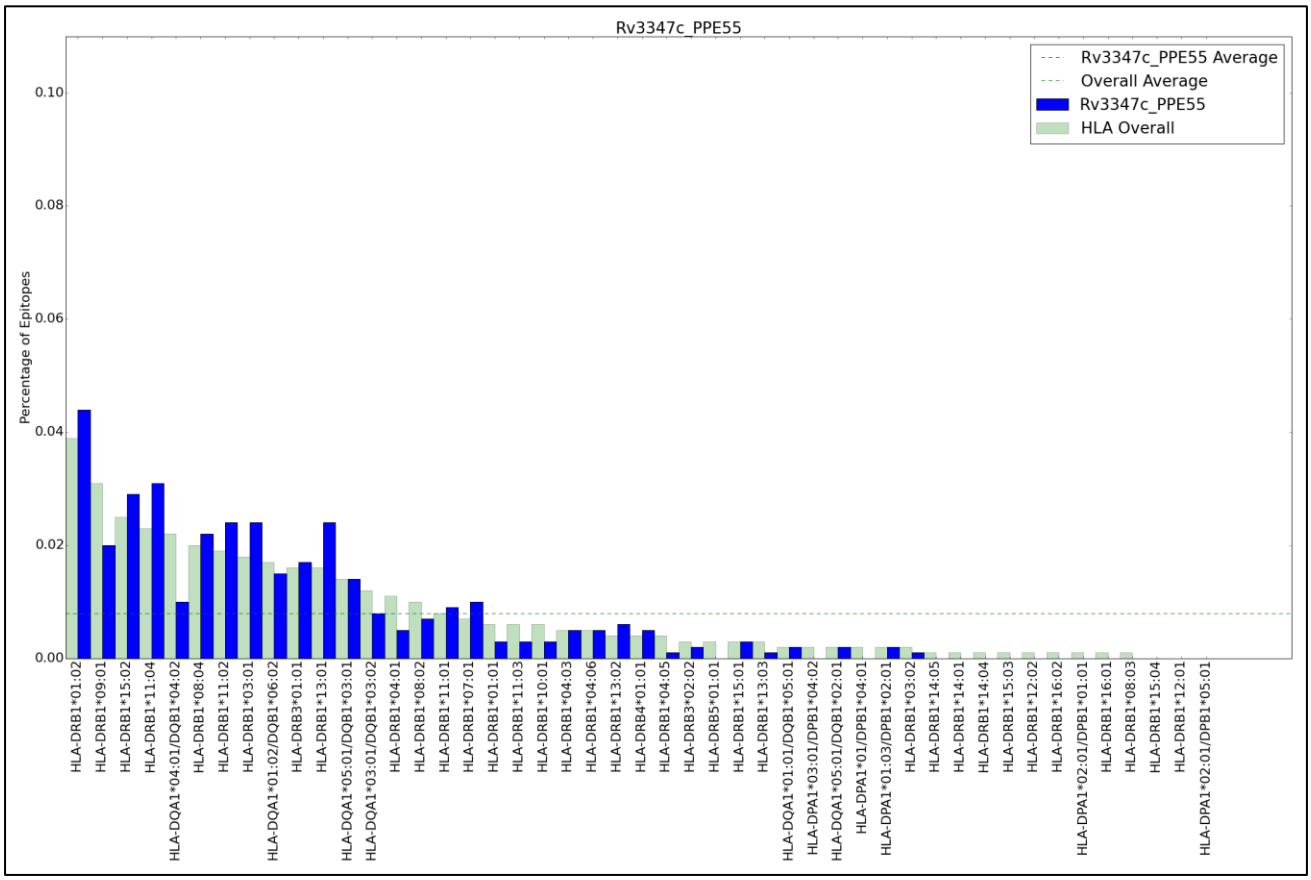


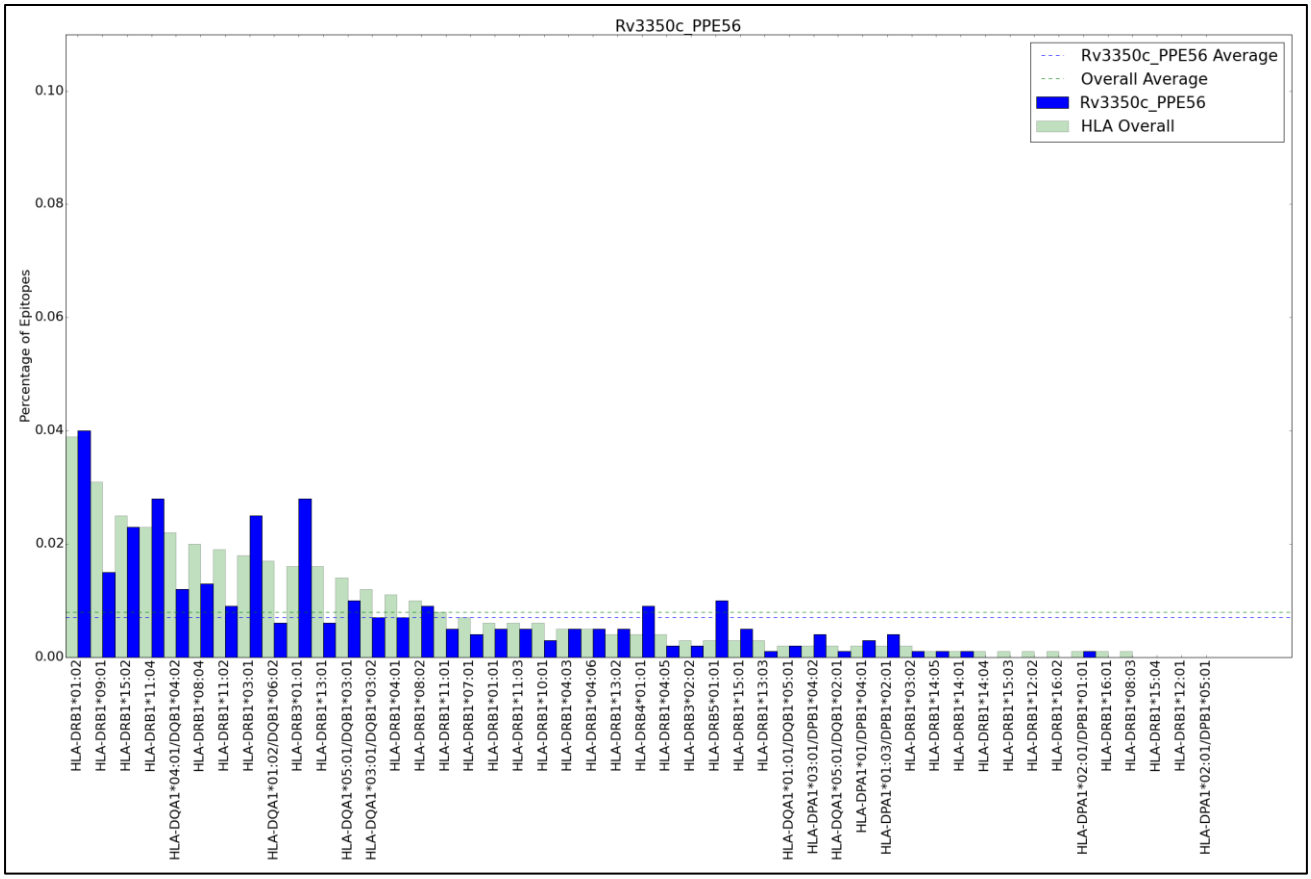


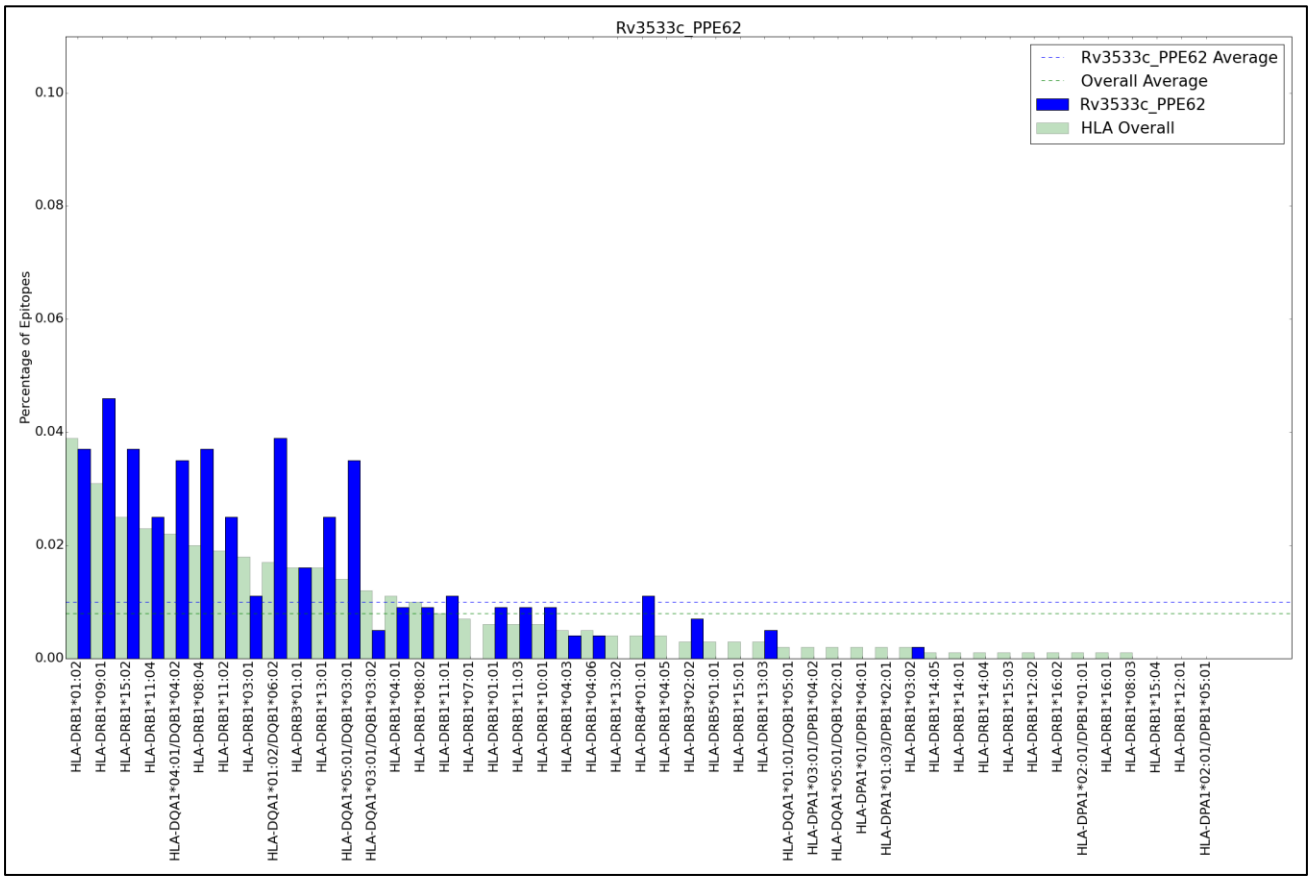


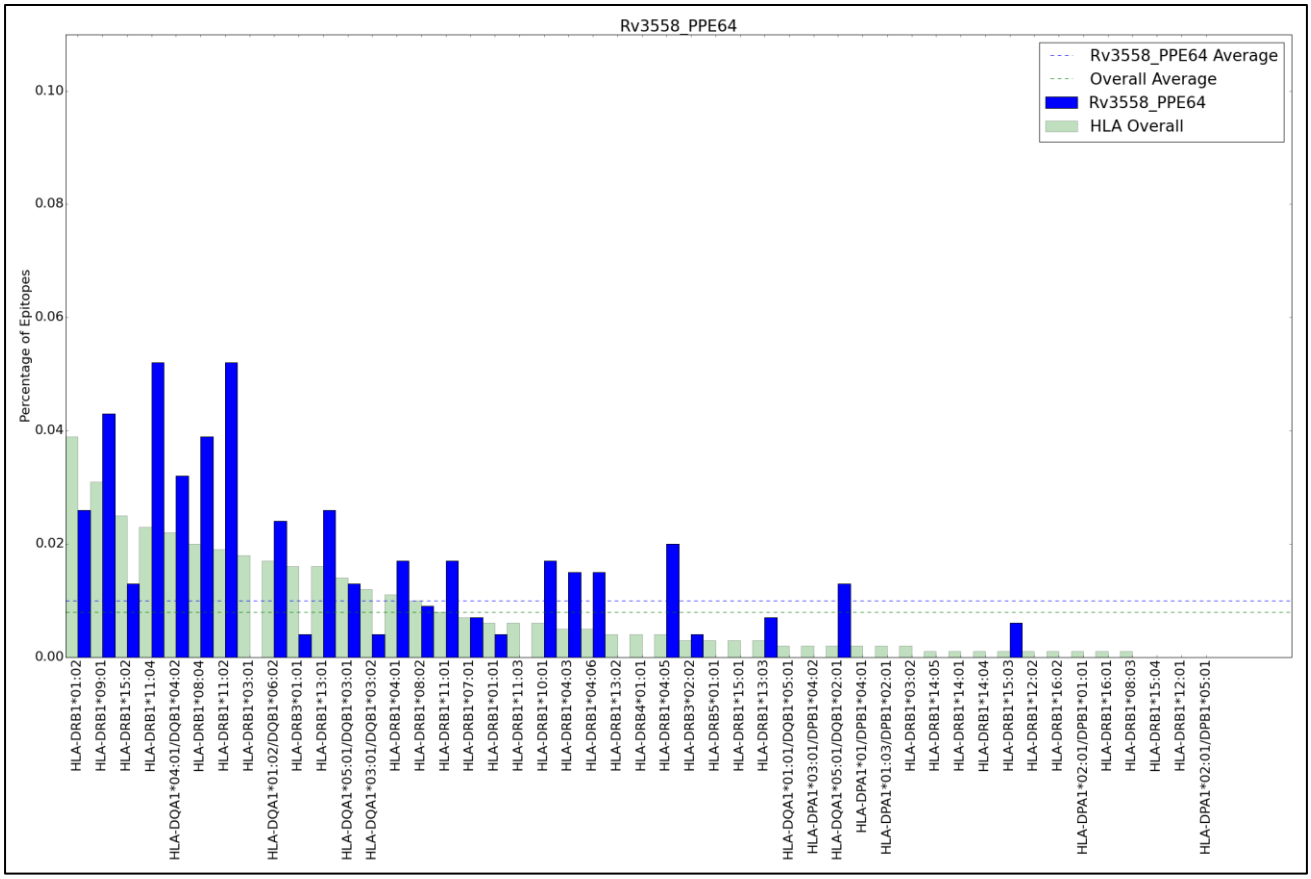












Chapter 5: Genetic Diversity of the PPE_MPTR Proteins

5.1 Introduction

5.1.1 PPE_MPTR genomics

The genes encoding the PE/PPE proteins make up approximately 10% of the *Mycobacterium tuberculosis* genome (Cole *et al*. 1998), of which the PPE_MPTR family is a subset consisting of 23 proteins (Chapter 1, Figure 1.1). The PPE_MPTR protein sequence includes the conserved PPE region of roughly 170-180 amino acids, and the MPTR region which ranges from between 420 to 3700 amino acids. The genes encoding the PPE_MPTR proteins are associated with imperfect repeats and are hotspots for recombination events and mutations (Cole *et al*. 1998; McEvoy *et al*. 2009). Given that *M. tuberculosis* is highly clonal, it has been hypothesized that recombination within the PPE proteins has been influential in the evolution of *M. tuberculosis* and a major contributor to genetic diversity (Liu *et al*. 2006).

5.1.2 Determining the genetic diversity of the PPE_MPTR proteins

The function of the PPE_MPTR proteins is largely unknown, and characterising the genetic diversity of these proteins is an essential step in improving our understanding of the possible role they may play in *M. tuberculosis* pathogenesis. Given the characteristics of the PPE_MPTR gene sequences, most specifically the imperfect repeats which characterise this family, as well as the limitations in the current sequencing technology, this has proven difficult. Repeated gene sequences cause technical challenges for sequence alignment and assembly programs given the short read length of most next generation sequencing platforms, as they create ambiguities in the alignment and assemblies (Treangen & Salzberg 2012). As a result, many genetic variation and phylogenic studies of the *M. tuberculosis* genome often exclude all of the PE/PPE proteins altogether. Even in an analysis focused specifically on the genetic diversity of the PE_PGRS protein family, which along with the PPE_MPTR proteins, are the most genetically diverse regions within the *M. tuberculosis* genome, polymorphic sites which included large indels (>25% of the gene) and frameshift indels were excluded from the analysis of nucleotide diversity (Copin *et al*. 2014). In a comparative analysis of all of the *M. tuberculosis pe/ppe* genes, it was shown that the majority of genetic diversity seen within the PPE_MPTR family is due to macro-mutations including homologous recombination, IS*6110* integration, partial and whole gene deletions rather than micro-mutations such as single nucleotide polymorphisms (SNP's) and small indels (McEvoy *et al*. 2012). Therefore following a similar approach of ignoring largely polymorphic regions within the PPE_MPTR proteins would result in a large portion of certain *ppe_mptr* genes discarded and the potential biological reason and importance for these macro-mutations overlooked.

Different methods for analysing and quantifying the diversity of a gene across multiple strains exist, including the approach used by Copin *et al*. (2014) for the *pe_pgrs* genes where nucleotide diversity (π) and indel diversity was calculated using the Nei method (Nei 1987). These and various other methods have been made available using software such as the Molecular Evolutionary Genetics Analysis (MEGA) toolkit (Tamura *et al*. 2011). However, these approaches require a multiple alignment of the various strains used in the analysis which can be prone to errors when investigating genes containing highly polymorphic regions such as those encoding the PE_PGRS and PPE_MPTR protein families, often necessitating the need to exclude these regions. IS*6110* insertions, recombination and large gene deletions within the *ppe_mptr* genes result in each genome containing a combination of unique lineage-specific segments, regions shared with a subset of other strains and segments conserved among all the genomes included (Darling *et al*. 2004), further complicating a multiple alignment. Certain members of the PPE_MPTR family have varying copy numbers of the tandem repeats between different strains (section 5.3.2 below), resulting in varying lengths for the same gene across different strains, which can also result in ambiguous multiple alignments. Typical genetic diversity scores such as nucleotide diversity (π) annotate the large differences in the size of genes between strains as large indels rather than a difference in the copy number of tandem repeats. IS*6110* insertion elements which have resulted from one insertion event are also counted as many nucleotide differences (Nei 1987), all of which overestimate the actual genetic diversity. A clustering approach, where repeats, and large indels are collapsed into single nucleotide markers has been applied to the human genome (Teixeira-Silva *et al*. 2013), and could similarly be applied to the *ppe_mptr* genes. Genetic diversity scores are also sensitive to the choice of strains to be included in the analysis making it extremely important to ensure that

a large number of strains from various lineages are included to make any quantitative score a reliable one. This also makes comparing the scores between studies which have used different strains unreliable. Given these observations, quantitative genetic diversity scoring methods may not be suitable for the *ppe_mptr* genes.

5.1.3 Current approach used

Given the challenges associated with determining the genetic diversity of the *ppe_mptr* genes, various approaches were used within this chapter to try and overcome these. This has included identifying and removing IS*6110* insertion elements (Section 5.2.4) and using the tandem repeat finder (TRF) tool which masks repeat regions (Section 5.2.5). A multiple alignment of genes where the regions that complicate the multiple alignment algorithm are masked has resulted in more accurate alignments. Actual gene sequences of the masked regions can subsequently be added back into the alignment. Small regions where alignments are ambiguous have been excluded, and small variants such as SNP's and indels have only been included if they are seen in at least 2 strains (Section 5.2.6).

The focus of this chapter is determining whether the hypothesis presented within this thesis of "genetic diversity within the PPE_MPTR proteins differentially modulating human immune response" is valid or not. Therefore genetic variants have been investigated within the context of epitope density. This chapter does not try to quantify the genetic diversity similarly to methods used within the MEGA toolkit, but rather whether a genetic variant observed has an impact on epitope density within that region, and whether epitope dense versus epitope void regions can be explained by genetic diversity within those regions.

5.2 Methodology

5.2.1 Methodology summary

Figure 5.1 shows a summary of the methodology used to determine the genetic variation within each of the *ppe_mptr* genes and the effects of the different variants on epitope prediction.

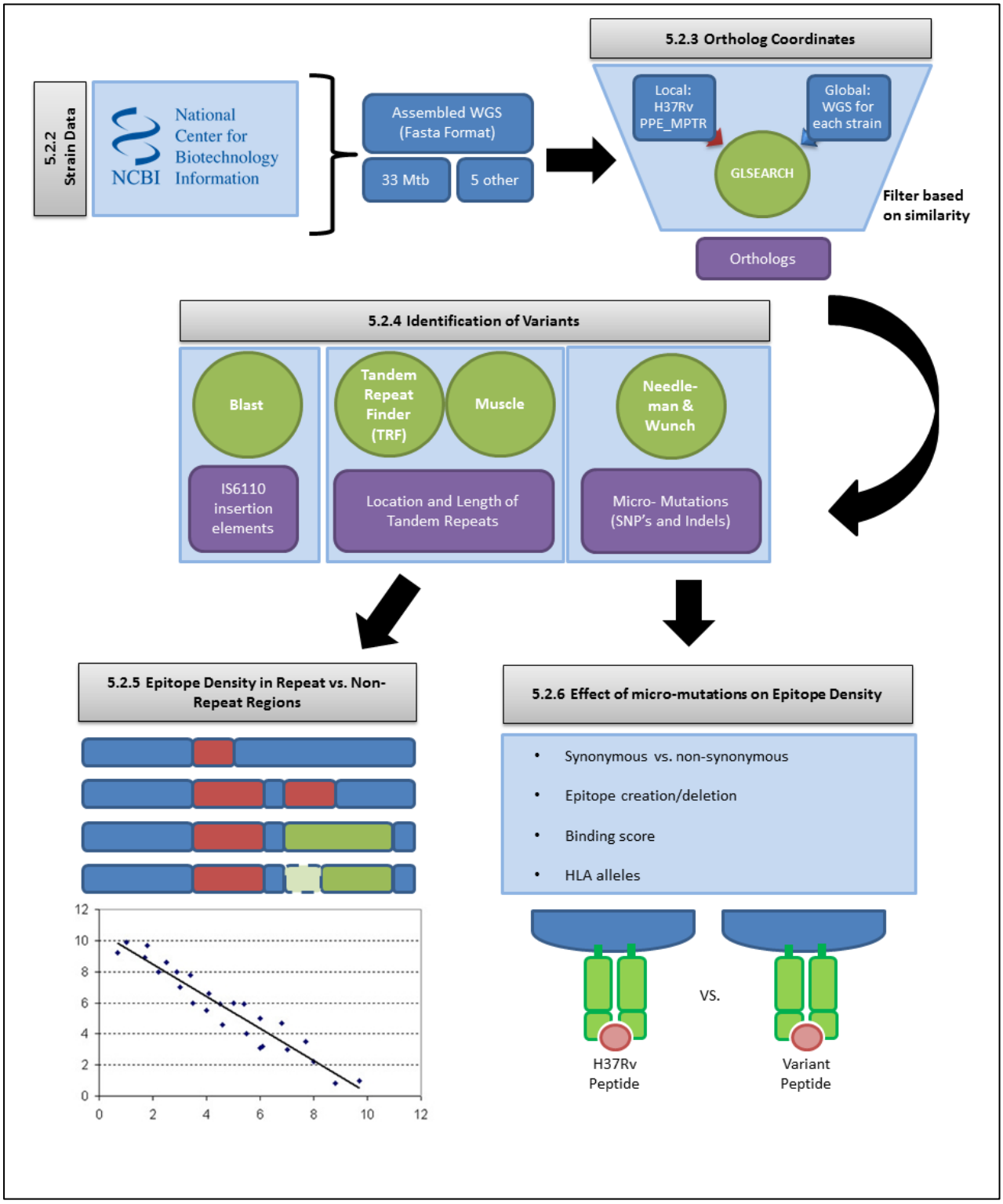


Figure 5.1: Genetic variation methodology. Methodology used to determine the genetic variation of the *ppe_mptr* genes involved using public WGS data from NCBI, identifying H37Rv orthologs in each strain and identifying IS*6110* insertion elements, tandem repeats and micro-mutations. The epitope density in repeat and non-repeat regions has been investigated as well as the effect of specific micro-mutations on epitope density.

5.2.2 Strain data

Publically available *M. tuberculosis* assembled whole genome sequences (WGS) were obtained from the National Center for Biotechnology Information (NCBI) website (http://www.ncbi.nlm.nih.gov/genome/genomes/166, accessed 29 September 2015), and TBDB (Galagan et al. 2010). Each sequence was downloaded as a complete FASTA sequence. (Gene annotations provided by the NCBI prokaryotic genome annotator were not used. See section 5.2.3.)

A total of 32 different *M. tuberculosis* WGS from various lineages were downloaded (Table 5.1). Information about each strain (such as the sequencing technology used and year submitted) was obtained from the metadata linked to each genome within the NCBI website where available. Lineage information for each strain was determined using CASTB, a publicly accessible web server which accepts WGS data and reports the results of spoligotyping, VNTR, lineage, and Beijing typing (Iwai *et al*. 2015).

In addition, five other mycobacteria*l* strains (*M. canetti, M. suricattae, M. bovis, M. africanum and M. bovis BCG)* were included in the analysis for comparative purposes. With the exception of *M. suricattae*, all other assembled WGS's were downloaded from the NCBI website in FASTA format (http://www.ncbi.nlm.nih.gov/genome/genomes/166, accessed 29 September 2015). The assembled WGS for *M. suricattae* was obtained from Dr. Anzaan Dippenaar from the SAMRC TB research unit, University of Stellenbosch.

The *M. tuberculosis* H37Rv strain was used as the reference strain. Gene sequences for the H37Rv *ppe_mptr* genes were obtained from TubercuList (Lew *et al*. 2011) in FASTA format. Similarly to the methodology followed during epitope prediction (section 4.2.1), three of the *ppe_mptr* genes (Rv0304c, Rv0354c and Rv2353) were not included in the analysis due to dissimilarity in protein sequence structure to the remaining PPE_MPTR proteins.

In total, including H37Rv, the additional 32 *M. tuberculosis* strains from NCBI, and 5 other mycobacteria, 38 strains were used to analyse the genetic variation within the *ppe_mptr* genes.

Table 5.2: Strain Information. Strain information including NCBI accession number, lineage, sequencing source and technology, and year of submission are given for each strain included in the analysis. Lineage was obtained from CasTB, while all other information was taken from supporting literature where available. For certain WGS, the sequencing technology used was not provided within the metadata linked to the genome. * Downloaded from TBDB

Strain Name	NCBI accession number	Lineage (CasTB)	Sequencing Source	Sequencing Technology	Year submitted
CDC1551	NC_002755.2	4	TIGR	Whole genome shotgun sequencing	2002
F11	NC_009565.1	4	Broad		2007
Haarlem	NC_022350.1	4	Broad		2006
KZN1435	NC_012943.1	4	Broad		2009
KZN4207	NC_016768.1	4	Broad		2009
KZN605	NC_018078.1	4	Broad		2009
W148	NZ_CP012090.1	2	Broad	Sanger, Illumina, PacBio.	2010
CTRI2	NC_017524.1	4	Research Institute for Physical-Chemical Medicine, Moscow, Russia		2013
CCDC5079	NC_021251.1	2	BGI		2011
CCDC5180	NC_017522.1	2	BGI		2011
719999	NC_020089.1	4	Center for Biotechnology, Bielefeld University		2012
Erdman ATCC 35801	NC_020559.1	4	National Center for Global Health and Medicine	ABI3730xl, Roche 454, Illumina GAIIx.	2012
NITR203	NC_021054.1	4	National Institute of Tuberculosis Research, Chennai, India	Illumina	2013
NITR206	NC_021194.1	4	National Institute of Tuberculosis Research, Chennai, India	Illumina	2013
EAI5	NC_021740.1	4	Department of Life Sciences, Mumbai University, Mumbai, India	Illumina	2013
HKBS1	NZ_CP002871.1	2	School of Biomedical Sciences, Faculty of Medicine, The Chinese University of Hong Kong		2011
BT2	NZ_CP002882.1	2	School of Biomedical Sciences, Faculty of Medicine, The Chinese University of Hong Kong		2011

Strain Name	NCBI accession number	Lineage (CasTB)	Sequencing Source	Sequencing Technology	Year submitted
BT1	NZ_CP002883.1	2	School of Biomedical Sciences, Faculty of Medicine, The Chinese University of Hong Kong		2011
K	NZ_CP007803.1	2	Seoul National University	Sanger, Illumina MiSeq	2015
KIT87190	NZ_CP007809.1	2	Korean Institute of Tuberculosis	Roche 454, GapFiller	2014
ZMC13264	NZ_CP009100.1		Affiliated Hospital of Zunyi Medical College	Ion Torrent	2014
ZMC1388	NZ_CP009101.1	4	Affiliated Hospital of Zunyi Medical College	Ion Torrent	2014
96075	NZ_CP009426.1		Northern Arizona University	Roche 454	2014
96121	NZ_CP009427.1	1	Northern Arizona University	Sanger dideoxy sequencing, Roche 454. Illumina	2014
4902	NZ_HG813240.1	2	JLU		2013
Kurono	NZ_AP014573.1	4	National Center for Global Health and Medicine	PacBio	2014
Beijing-like	NZ_CP010873.1	2	Universidad Nacional de Colombia - Sede Bogota	PacBio	2015
NITR204	CP005386.1	4	National Institute of Tuberculosis Research, Chennai, India	Illumina	2013
SCAID187	NZ_CP012506.2	2	SCAID	Illumina Hiseq 2000, Ion Torrent	2015
RGTB327	*	4	Rajiv Gandhi Centre for Biotechnology	Roche 454	2012
RGTB432	*	4	Rajiv Gandhi Centre for Biotechnology	Roche 454	2012
C	*	4	Broad	Whole genome shotgun sequencing	2005

.

5.2.3 Ortholog coordinates

The coordinates of the H37Rv *ppe_mptr* orthologs from each strain were needed in order to determine genetic variation between strains within each gene. The prokaryotic genome annotations provided by the NCBI are predicted genes regions and have not been curated (Tatusova *et al*. 2016). Visual inspection of pairwise alignments between the H37Rv *ppe_mptr* genes and the orthologs given by the prokaryotic genome annotations showed poor correlations and missing sections at the beginning and/or end of certain gene sequences (data not shown). It was therefore decided that these automated annotations would not be used.

Coordinates of the H37Rv *ppe_mptr* orthologs were obtained using the GLSEARCH command line tool (version 36.3.8) to perform a global: local alignment between each complete FASTA sequence for each strain and each H37Rv *ppe_mptr* gene sequence. The output of the GLSEARCH tool is in the format of a list of possible alignments with a corresponding similarity score. Similarity scores were used to filter alignments based on the following criteria:

- Greater than 90% similarity scores: Coordinates given by GLSEARCH used as is.
- Between 70% and 90% similarity: A pairwise alignment between the H37Rv gene (extended by 500 nucleotides on each side) and the possible ortholog (extended by 500 nucleotides on each side) was performed using the Needleman and Wunsch algorithm (Needleman & Wunsch 1970). Alignments were visually inspected to determine correct coordinates or discard the alignment.
- Less than 70% similarity scores: Not used.

For each strain, the coordinates of each gene were checked for overlap (i.e. no two genes had overlapping coordinates within one strain). Where overlapping coordinates were found, Needleman and Wunsch pairwise alignments (Needleman & Wunsch 1970) were visually inspected, and orthologs discarded where necessary. In many cases an overlap between Rv3343c, Rv3347c and Rv3350c was found. After visual inspection of the Needleman and Wunsch pairwise alignments for these genes, the coordinates for Rv3343c were not able to be confirmed for many of the strains. High sequence variation across strains was also observed for Rv3347c and Rv3350c. The TB Database (TBDB) (Galagan *et al*. 2010), which also provides WGS data for various *M. tuberculosis* strains, has provided a list of orthologs of all of the genes within each strain available in their database. Similarly, this list does not differentiate between Rv3343c, Rv3350c and Rv3347c, but rather lists them as one set for H37Rv corresponding to another set of one or two genes within other strains. It could be hypothesized that these genes have resulted from duplication events, with a varying number and/or combination present in different strains. Given the high similarity of all three of these genes within H37Rv, it is therefore extremely difficult to distinguish whether the correct ortholog for each gene was found within each strain, and the high sequence variation observed could be a result of the incorrect ortholog being used. Therefore genetic variation analysis for Rv3343c, Rv3350c and Rv3347c between strains was not performed. Identification of IS*6110* insertion elements (section 5.2.4) and tandem repeats (5.2.5) was therefore restricted to the gene sequence for H37Rv for these 3 genes.

For certain strains it was not possible to find the coordinates of all of the remaining 17 *ppe_mptr* genes. Missing coordinates may either be due to poor sequencing or missing reads within these regions, or an actual deletion of certain *ppe_mptr* genes within these strains. The number of strains used to determine genetic variation within each of the 17 *ppe_mptr* genes is shown in Table 5.2.

Table 5.3: Number of strains used to determine genetic variation. For certain genes, the ortholog coordinates within each strain were not found and therefore not all 38 strains were included.

	Gene	Number of Strains		Gene	Number of Strains
1.	Rv0305c	36	**10.**	Rv1917c	34
2.	Rv0355c	37	**11.**	Rv1918c	37
3.	Rv0442c	36	**12.**	Rv2356c	31
4.	Rv0755c	37	**13.**	Rv2608	37
5.	Rv0878c	37	**14.**	Rv3144c	36
6.	Rv1135c	36	**15.**	Rv3159c	37
7.	Rv1548c	37	**16.**	Rv3533c	38
8.	Rv1753c	37	**17.**	Rv3558	37
9.	Rv1800	35			

5.2.4 IS*6110* insertion elements

The GLSEARCH command line tool was used to perform a global: local alignment (using default parameters) between each of the *ppe_mptr* orthologs from each strain and the DNA sequence for an IS*6110* insertion element (provided by Dr. Anzaan Dippenaar from the SAMRC TB research unit, University of Stellenbosch). Similarity scores of at least 99% were used. Once identified, IS*6110* insertions were removed from the gene sequence for further analysis of genetic variation.

5.2.5 Tandem repeats

The Tandem Repeat Finder (TRF) tool (Benson 1999) was used to find the presence and number of tandem repeats within each of the *ppe_mptr* genes for all strains. This tool also allows the user to output a masked version of the gene sequence where areas containing tandem repeats are masked with an 'X'. A multiple alignment of the masked sequences for each gene was performed using Muscle (Edgar 2004). Subsequently, the original gene sequences of the masked regions were reintroduced into the multiple alignments.

When performing the identification of tandem repeats, two broad categories of tandem repeats were identified. The typical MPTR tandem repeats as well as longer tandem repeats (Non-MPTR). Repeats were separated into each category based on length. MPTR repeats were in multiples of 15 bp, while longer repeats where between 70-90 bp. Alignment and clustering of repeated segments was used to confirm the category of repeat regions. For each *ppe_mptr* gene, the number of repeats was compared across strains to determine whether the copy number was constant or variable.

5.2.6 Determination of variants

In addition to the multiple sequence alignment described in section 5.2.5, for each of the 17 *ppe_mptr* genes, a Needleman and Wunsch pairwise alignment (Needleman & Wunsch 1970) between the H37Rv gene sequence and its corresponding ortholog from each strain was performed using the needle command line tool from EMBOSS. The gap opening penalty was set to 100 and gap extension penalty was set to 1 in order to favour alignments with fewer long gaps over alignments with many short gaps given the large indels and variable copy number of repeats within the *ppe_mptr* genes. Pairwise alignments were used to determine genetic variants including SNPs, insertions and deletions. Each variant found was compared to variation results obtained using the multiple sequence alignment described in section 5.2.5. Differences were visually inspected to determine the reason for the difference and which variant (if any) was correct.

Due to the known high sequencing error rate within the *pe/ppe* genes, only variants found in at least 2 strains compared to H37Rv were confirmed as variants. Furthermore, areas where the alignment with the H37Rv gene was ambiguous (i.e. an insertion or deletion could be placed in more than one area and therefore a different set of resulting SNP's reported) were discarded from further analysis. Insertions and deletions in areas with repeating bases were also excluded due to the high probability that these indels may be due to sequencing error rather than true genetic variation.

Rv1917c is one of the *ppe_mptr* genes that is extremely variable, particularly in the length of the gene between strains. It contains IS*6110* insertion elements and both MPTR repeats as well as the longer repeat regions in different copy numbers across strains. Removing IS*6110* insertion elements and masking repeat regions as described in section 5.2.4 and 5.2.5 above still results in an ambiguous multiple alignment, which can result in a different set of SNP's called depending on which alignment is chosen. This may be due to additional partial gene deletions within certain strains. For this reason, micro-mutations were not investigated for Rv1917c.

5.2.7 Effect of micro-mutations on epitope prediction

Micro-mutations are defined as SNP's and small indels (less than 9 bp in length). The effect of each micro-mutation on the protein and epitope density within the area of the mutation was investigated. For each SNP identified, the effect on epitope density within that region was investigated using the following approach:

- Determining the effect of the SNP on the protein sequence (i.e. distinguishing between a synonymous SNP (sSNP) versus non-synonymous SNP (nsSNP). Variant gene sequences (containing the specific SNP) were transcribed and translated into a protein sequence using the Biopython package (Cock *et al*. 2009), and aligned to the H37Rv protein sequence to determine the position and amino acid change.

- Re-performing epitope prediction with variant protein sequence. When epitope prediction is preformed, the protein sequence is segmented into overlapping peptides of 15 amino acids in length. A non-synonymous SNP can therefore affect 15 different peptides. For each of these variant peptides, epitope prediction was rerun with each of the 47 HLA alleles included in Chapter 4.

- Comparison of results from epitope prediction using H37Rv and the variant protein sequence. Prediction results were analysed to determine whether the genetic variant has resulted in epitope creation or deletion, effect on average predicted binding score and whether the number of HLA alleles the peptide is able to bind to has changed.

5.2.8 Epitope density within repeat versus non-repeat regions

Epitope density along the length of the PPE_MPTR proteins shows patterns of fluctuation (Chapter 4, Section 4.3.2). Whether these fluctuating patterns correlate with repeat versus non-repeat areas was investigated for H37Rv using two approaches:

- The total percentage of repeat regions (i.e. regions masked by the Tandem Repeat Finder) out of the length of the gene was compared to overall epitope density for each PPE_MPTR protein.

- Repeat and non-repeat regions alternate along the length of the PPE_MPTR proteins (An example graphical representation is shown in Figure 5.2). Epitope density within each repeat and non-repeat region was calculated separately, and the distribution of epitope density in repeat versus non-repeat regions was compared overall for the PPE_MPTR proteins as well as separately for each protein individually. Coordinates of repeat regions within each gene was converted into protein coordinates which were used to determine epitope density within each individual region.

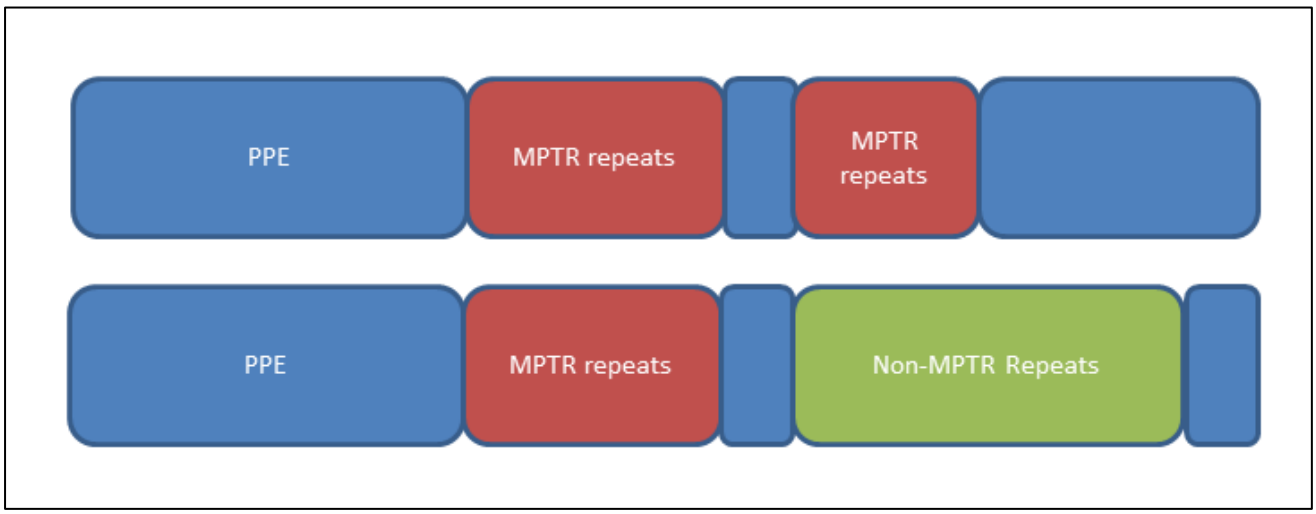


Figure 5.2: Example graphical representation of alternating repeat and non-repeat regions. MPTR repeats (red) as well as longer non-MPTR repeats (green) can be found within the *ppe_mptr* genes dispersed between non-repeated gene regions (blue). The copy number of both types of repeats can vary between strains changing the length of the gene sequence and resulting protein.

5.3 Results

5.3.1 Summary of genetic variation

Table 5.3 shows a summary of the type and amount of genetic variation seen within each *ppe_mptr* gene sequence.

Table 5.4: Summary of genetic variation observed. For each gene, the number of sSNP, nsSNP's and small indels is shown, as well as whether IS*6110* insertion elements were found within at least one strain. The presence of MPTR and long non-MPTR repeats is shown as well as whether the copy number was found to be variable or constant. An acceptable multiple alignment of Rv1917c was not found and therefore micro-mutations for Rv1917c are not available (section 5.2.6). Ortholog coordinates for Rv3343c, Rv3347c, and Rv3350c could not be confirmed and therefore no micro-mutations between strains could be identified (section 5.2.3). IS*6110* insertions and tandem repeats for these genes were searched for using only H37Rv, and therefore copy number of repeats could not be calculated.

Gene	sSNP's	nsSNP's	Small Indels	IS*6110* insertion	Imperfect Tandem Repeats (MPTR)	Copy Number	Long Imperfect Tandem Repeats (non-MPTR)	Copy Number
Rv0305c	2		1*		Yes	Variable		
Rv0355c	5	8	1		Yes	Variable		
Rv0442c		1			Yes	Constant		
Rv0755c		1			Yes	Constant		
Rv0878c	1				Yes	Constant		
Rv1135c				Yes	Yes	Constant		
Rv1548c		1			Yes	Constant		
Rv1753c	2	2			Yes	Variable	Yes	Variable
Rv1800		4		Yes	***			
Rv1917c				Yes	Yes	Variable	Yes	Variable
Rv1918c		4	1		Yes	Constant	Yes	Constant
Rv2356c		1			Yes	Constant		
Rv2608		1			Yes	Constant		
Rv3144c	1	2			Yes	Constant		
Rv3159c		5	1**		Yes	Constant		
Rv3343c					Yes			
Rv3347c					Yes			
Rv3350c					Yes			
Rv3533c		1			Yes	Constant		
Rv3558	1	3	1**		Yes	Constant		
TOTAL	12	34	5					

*H37Rv not wild type

** Within repeated area and may therefore be due to sequencing error (low confidence).

***no MPTR repeats were identified by the TRF tool for Rv1800.

5.3.2 IS*6110* insertion elements

IS*6110* insertion elements were found within three of the *ppe_mptr* genes, Rv1153c, Rv1800, and Rv1917c. Table 5.4 shows the position of the IS*6110* elements and which strains they were found in.

Table 5.5: Genes containing IS*6110* insertion elements. Position of the IS*6110* insertion element within the gene is shown along with the strains containing the insertion.

Gene	Position	Strains
Rv1153c	1150 - 2524	4902. 96075, HKBS1, BT1, CCDC5180, BT2, Beijing_like, CCDC5079
Rv1800	1770 - 3142	CTRI2
	976 – 2350	CCDC5180
Rv1917c	1587 - 2964	KIT87190, BT1, HKBS1, K, W148, Beijing_like
	1139 - 2511	96075
	1451 - 2826	4902

For Rv1800, the IS*6110* insertion element was found in two strains but in two distinct places within the genome. These regions did not overlap. For Rv1917c, the position of the IS*6110* insertion elements between different strains all overlapped with each other and the difference in position was due to other variation within the genomes of these strains (either variable copy numbers of repeats, or other insertions or deletions) before the IS*6110* insertion element.

5.3.3 Effect of micro-mutations on epitope prediction

Table 5.5 shows the small indels identified within each *ppe_mptr* gene.

Table 5.6: Small indels identified within the *ppe_mptr* genes. Information for each indel identified and the consequence of the variant on the protein is given. Indels for Rv1917c could not be identified determined given the ambiguity in the multiple alignment across strains for this gene. Ortholog coordinates for Rv3343c, Rv3347c, and Rv3350c could not be confirmed and therefore indels were not identified for these genes. No indels were identified for Rv0442, Rv0755c, Rv0878c, Rv1135c, Rv1548c, Rv1753c, Rv1800, Rv2356c, Rv2608, Rv3144c, and Rv3533c.

Protein	Position	Size	H37Rv Allele	Alt Allele	% of strains	Strains containing Variant	Consequence on Protein
Rv0305c	2428	1	T	-	100% (37/37)	All except H37Rv	Protein sequence changed after 809 amino acids.
Rv0355c	9888	2	-	AT	24.3% (9/37)	RGTB432, 96121, *M. canetti, M. africanum, M. suricattae, M. bovis*, BCG, NITR206, EAI5	Protein sequence changed after 3296 amino acids, only affecting the Last 4 amino acids. However not a multiple of 3, and stop codon at the end of the protein is disrupted.
Rv1918c	1797	1	-	A	5.4% (2/37)	*M. bovis*, BCG	Change in protein sequence from amino acid 600 – 625, followed by a stop codon and therefore a truncated protein.
Rv3159c	178	3	GCG	-	37.8% (14/37)	BCG, *M. bovis, M. suricattae*, BT2, *M. africanum*, CCDC5180, Beijing-Like, BT1, KIT87190, K, CCDC5079, 96075, W148, 4902	Deletion of 1 amino acid. No change to the rest of the protein. This deletion is within a section with repeating "GCG". Therefore strains may have a variable number of this codon. It may also however be a sequencing error and not a real deletion.
Rv3558	81	3	GGC	-	13.5% (5/37)	KZN1435, KZN605, KZN4207, F11, CTRI2	Deletion of 1 amino acid. No change to the rest of the protein. This deletion is within a section with repeating "GGC". Therefore strains may have a variable number of this codon. It may also however be a sequencing error and not a real deletion.

Indels for Rv3159c and Rv3558 may be due to sequencing error as they are found within regions containing a repeated codon. Effect on epitope prediction for these variants was therefore not investigated. The indel within Rv1918c resulted in a truncated protein, and therefore the effect on epitope prediction was not investigated. Given that the indel found within Rv0355c only affected the last 4 amino acids, the effect on epitope prediction was also not investigated.

A single nucleotide deletion was found in Rv0305c for all strains when compared to H37Rv. H37Rv protein sequence is therefore not the wild type sequence. From position 809 in the protein, the amino acid sequence is different for H37Rv compared to all other strains. This could potentially have a large impact on the results of epitope prediction for this protein. Epitope prediction was therefore rerun using a representative "wild-type" sequence from among the strains against all 47 HLA alleles used in Chapter 4, and compared to the results using H37Rv. The protein sequence changes within the MPTR region and there results are shown for epitopes found within the MPTR region only (Table 5.6).

Table 5.6: Comparison between epitope prediction results for Rv0305c in H37Rv and "wild-type" sequence. The number of epitopes within the MPTR region, average binding score and average number of alleles has been compared.

	H37Rv	"wild-type"
Number of MPTR Epitopes	230	244
Average binding score	0.456	0.522
Average number of HLA alleles	2.14	1.81

The number of epitopes within the MPTR region increased from 230 to 244 epitopes. Protein length remained the same and therefore epitope density has increased by 6.1%. The average binding score of epitopes also increased, however the average number of alleles each epitope is able to bind to decreased.

Table 5.7 shows the SNP's identified within each *ppe_mptr* gene.

Table 5.7: SNP's identified within the *ppe_mptr* genes. Information for each SNP including which strains contained the variant and the effect on the protein is shown. Micro-mutations for Rv1917c could not be identified given the ambiguity in the multiple alignment across strains for this gene. Ortholog coordinates for Rv3343c, Rv3347c, and Rv3350c could not be confirmed and therefore micro-mutations were not identified for these genes.

Protein	Position	H37Rv Allele	Alternate Allele	% of strains	Strains containing Variant	Type	Consequence on Protein
Rv0305c	1359	T	C	33.3% (12/36)	HKBS1, CCDC5180, Beijing-Like, BT1, BT2, ZMC1388, CCDC5079, 96075, NITR203, W148, SCAID187, 4902	sSNP	None
	2799*	T	G	80.6% (29/36)	All except RGTB327, Kurono, F11, KZN4207, KZN605, KZN1435, CTRI2	sSNP	None
Rv0355c	353	T	C	5.4% (2/37)	RGTB432, 96121	nsSNP	V118A
	2144	G	A	32.4% (12/37)	Beijing-Like, ZMC1388, NITR203, SCAID187, CCDC5180, BT2, CCDC5180, BT2, CCDC5079, 9605, W148, HKBS1, 4902, BT1	nsSNP	G715D
	3357	C	A	5.4% (2/37)	*M. africanum, M. suricattae*	sSNP	None
	3578	C	G	5.4% (2/37)	NITR206, EAI5	nsSNP	A1193G
	3924	C	T	8.1% (3/37)	Erdman, 719999, Haarlem	sSNP	None
	4348	T	C	5.4% (2/37)	KIT87190, K	nsSNP	F1450L
	5840	G	A	8.1% (3/37)	Erdman, 719999, Haarlem	nsSNP	G1947D
	5982*	G	A	67.6% (25/37)	All except RGTB327, C, Erdman, Kurono, F11, KZN4207, KZN605, KZN1435, 719999, CDC1551, CTRI2, Haarlem	sSNP	None
	7209	A	C	8.1% (3/37)	*M. canetti, M. bovis*, BCG	sSNP	None
	7912	G	A	10.8% (4/37)	RGTB432, 96121, NITR206, EAI5	nsSNP	G2638S
	8022	G	T	8.1% (3/37)	NITR204, NITR203, NITR206	nsSNP	L2674F
	9699	C	T	40.5% (15/37)	Beijing-Like, ZMC1388, ZMC13264, NITR203, SCAID187, BT2, CCDC5180, CCDC5079, 96075, W148, HKBS1, 4902, BT1, KIT87190, K	sSNP	None
	9748	A	T	24.3% (9/37)	RGTB432, 96121, *M. canetti, M. africanum, M. suricattae, M. bovis*, BCG, NITR206, EAI5	nsSNP	I3250F
Rv0442c	23	G	A	5.6% (2/36)	BCG, M. bovis	nonsense	Premature stop codon.
Rv0755c	1634	G	A	21.6% (8/37)	RGTB432, *M. africanum*, BCG, *M. suricattae, M. bovis*, 96121, EAI5, *M. canetti*	nsSNP	R545K
Rv0878c	1008	C	T	10.8% (4/37)	RGTB432, 96121, NITR206, EAI5	sSNP	None
Rv1135c	None						
Rv1548c	773*	A	G	97.3% (36/37)	All except Kurono	nsSNP	D258G
Rv1753c	1459	C	T	5.4% (2/37)	ZMC13264, C	sSNP	None
	1463*	A	C	91.9% (34/37)	All except Kurono, ZMC13264, C	nsSNP	N488T

Protein	Position	H37Rv Allele	Alternate Allele	% of strains	Strains containing Variant	Type	Consequence on Protein
	1775	A	C	5.4% (2/37)	ZMC13264, ZMC1388	nsSNP	N592T
	1875	C	A	21.6% (8/37)	C, Haarlem, *M. africanum, M. suricattae, M. bovis, Erdman*, 719999, BCG	sSNP	None
Protein	Position	H37Rv Allele	Alternate Allele	% of strains	Strains containing Variant	Type	Consequence on Protein
Rv1800	432	T	G	17.1% (6/35)	BCG, *M. bovis, M. africanum*, 96121, RGT432, EAI5	nsSNP	C144W
	449	C	T	14.3% (5/35)	C, Erdman, 719999, CDC1551, Haarlem	nsSNP	A150V
	757	T	G	8.6% (3/35)	BCG, *M. bovis, M. africanum*	nsSNP	F253V
	1508	T	C	5.7% (2/35)	BCG, M. bovis	nsSNP	V503A
Rv1918c	1871	G	A	5.4% (2/37)	C, RGTB432	nsSNP	G624D
	2602	A	C	5.4% (2/37)	BCG, *M. bovis*	nsSNP	S868R
	2630	G	A	5.4% (2/37)	NITR206, EAI5	nsSNP	G877D
	2687*	T	C	67.6% (25/37)	All except RGTB327, C, Erdman, Kurono, F11, KZN4207, KZN605, KZN1435, 719999, CDC1551, CTRI2, Haarlem	nsSNP	L896S
Rv2356c	539	C	T	16.1% (5/31)	NITR203, NITR204, NITR206, ZMC13264, EAI	nsSNP	S180L
Rv2608	841	C	T	10.8% (4/37)	*M. suricattae; M. Bovis; M. africanum*; BCG	nsSNP	P281S
Rv3144c	676	G	A	8.3% (3/36)	NITR204, NITR206, EAI5	nsSNP	G226S
	996	G	A	5.6% (2/36)	ZMC1388, ZMC13264	sSNP	None
	1198*	A	C	69.4% (25/36)	All except RGTB327, C, Erdman, Kurono, F11, KZN4207, KZN605, KZN1435, 719999, CDC1551, CTRI2, Haarlem	nsSNP	K400Q
Rv3159c	97*	C	G	59.5% (22/37)	All except NITR204, BCG, *M. bovis, M. suricattae*, RGTB327, Kurono, F11, KZN4207, KZN605, KZN1435, *M. africanum*, CTRI2, NITR206, EAI5	nsSNP	R33G
	612	C	A	5.4% (2/37)	BCG, M. bovis	nsSNP	D204E
	1013	G	T	5.4% (2/37)	NITR206, EAI5	nsSNP	G338V
	1190	C	T	5.4% (2/37)	NITR206, EAI5	nsSNP	T397I
	1681	A	G	5.4% (2/37)	NITR206, EAI5	nsSNP	T561A
Rv3533c	1026	C	A	8.1% (3/37)	Erdman, 719999, Haarlem	nsSNP	S342R
Rv3558	634	C	T	8.1% (3/37)	KZN1453, KZN605, KZN4207	nsSNP	L212F
	916	G	A	10.8% (4/37)	96121, NITR206, EAI5, RGTB432	nsSNP	G306S

When epitope prediction is preformed, the protein sequence is segmented into overlapping peptides of 15 amino acids in length. A non-synonymous SNP can therefore affect 15 different peptides. For each nsSNP, epitope prediction was rerun for the 15 individual peptides affected by the amino acid change against all 47 HLA alleles used in chapter 4. For each peptide, the following was investigated:

- Whether the variant resulted in epitope destruction or creation.
- Whether the overall binding score changed by more than 10%.
- Whether the number and type of HLA alleles the peptide is able to bind to changed.

These results can be found in Appendix A which has been attached as an excel spread sheet given the large size of the table.

In certain cases, a variant was found in an epitope void area, and the change in amino acid did not result in epitope creation. The nsSNP's in these cases therefore had no effect on epitope density. This was seen for certain nsSNP's within Rv0355c, Rv1800, Rv1918c, Rv3159c, and Rv3558c; and for all SNP's within Rv0755c, Rv1548c, Rv1753, and Rv3533c. Certain variants found in epitope dense areas also did not affect the epitope density within those areas, i.e. the variant peptide was still a predicted epitope, with no change to binding score and number or type of HLA alleles. Certain other nsSNP's did result in a change in the overall binding score of the peptide. This is true for selected SNP's within Rv0355c, Rv1800, Rv1918c, Rv2608, Rv3144c, and Rv3159c. Particular nsSNP's also changed the type and number of HLA alleles the peptide was able to bind to. This can be seen within Rv0355c, Rv1800, Rv1918c, Rv2356c, Rv2608, Rv3144c, Rv3159c, and Rv3559). There were limited instances where epitopes were created within an epitope void area as a result of the variant. This was seen within Rv1918c, Rv2356c, Rv3159c, and Rv3558. In certain other cases, the nsSNP resulted in epitope deletion. This was seen in Rv0355c, Rv1800, Rv1918c, Rv3144, and Rv3159.

Overall, certain SNP's affected epitope prediction while others did not. A clear pattern of micro-mutations affecting HLA class II binding ability could not be identified.

5.3.4 Epitope density within repeat versus non-repeat regions

The correlation between regions containing tandem repeats and epitope density was investigated in two ways:

- The overall percentage of repeat regions within each gene sequence was compared to the overall epitope density within a protein (Figure 5.4)

- Given that repeat and non-repeat regions alternate along the length of the PPE_MPTR proteins (Figure 5.2), epitope density within each repeat and non-repeat region was calculated separately, and the distribution of epitope density within the different regions was compared overall for the PPE_MPTR proteins (Figure 5.5) as well as separately for each protein individually (Figure 5.6)

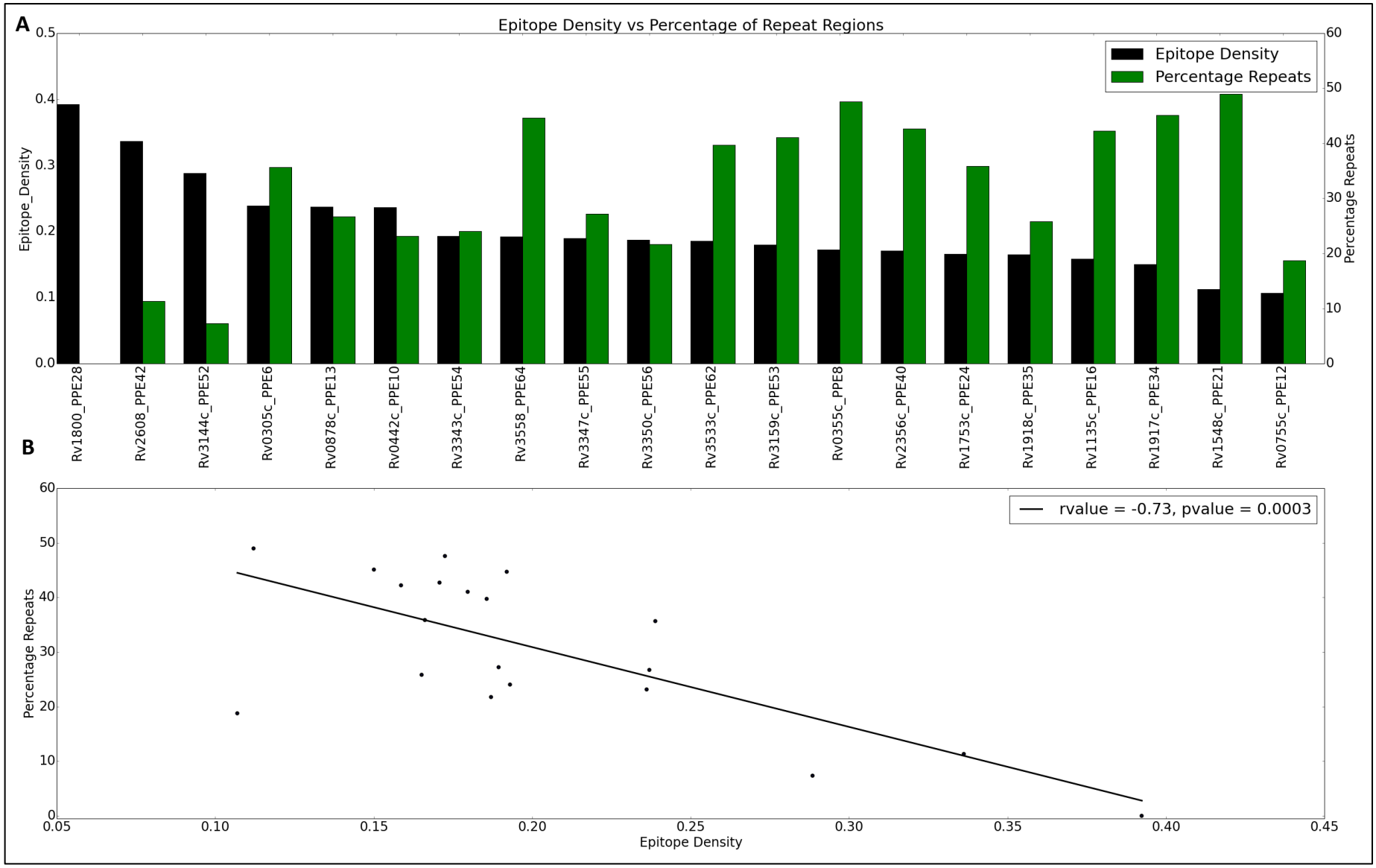


Figure 5.4: Comparison between epitope density and overall percentage of repeat regions. A) Epitope density is shown in black and percentage of repeats in shown in green. Proteins along the x-axis have been ordered in descending epitope density. B) Epitope density and percentage of repeats show a strong inverse correlation ($r = -0.73$, $p < 0.001$).

A strong inverse relationship between epitope density and percentage of repeats can be seen, indicating that those PPE_MPTR proteins with fewer repeat areas as a percentage of the total protein length have on average a higher epitope density.

Figure 5.5 shows the results when calculating the epitope density within each separate alternating repeat and non-repeat region along the length of all PPE_MPTR proteins.

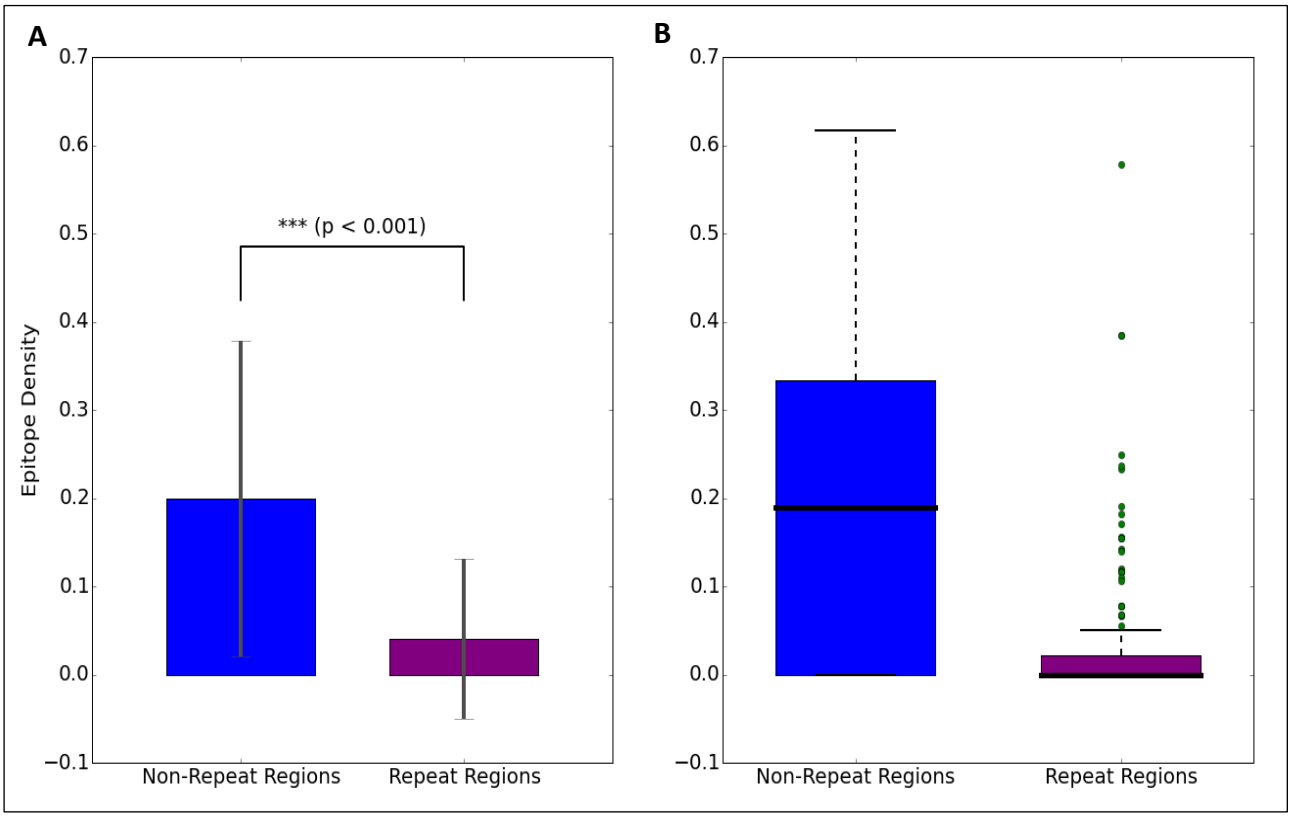


Figure 5.5 Comparison of epitope density within repeat and non-repeat regions across all PPE_MPTR proteins. A) Mean epitope density within non-repeated regions (blue) is compared to repeat regions (purple). Error bars show standard deviation in epitope density. Difference in mean epitope density between repeat and non-repeat areas is statistically significant ($p < 0.001$). B) Box plot showing the distribution of epitope density in non-repeat regions (blue) versus repeat regions (purple), with outliers shown in green.

Overall epitope density within the PPE_MPTR proteins is significantly higher within non-repeat areas, which suggests that the fluctuating patterns of epitope density along the length of the PPE_MPTR proteins is due to the alternating repeat and non-repeat regions. However within the repeat regions, outliers can be seen (Figure 5.5 B), which indicates that there are certain repeat regions with a higher epitope density comparable with the non-repeat regions. To investigate whether this is specific to certain PPE_MPTR proteins, we viewed this data separately for each PPE_MPTR protein (Figure 5.6).

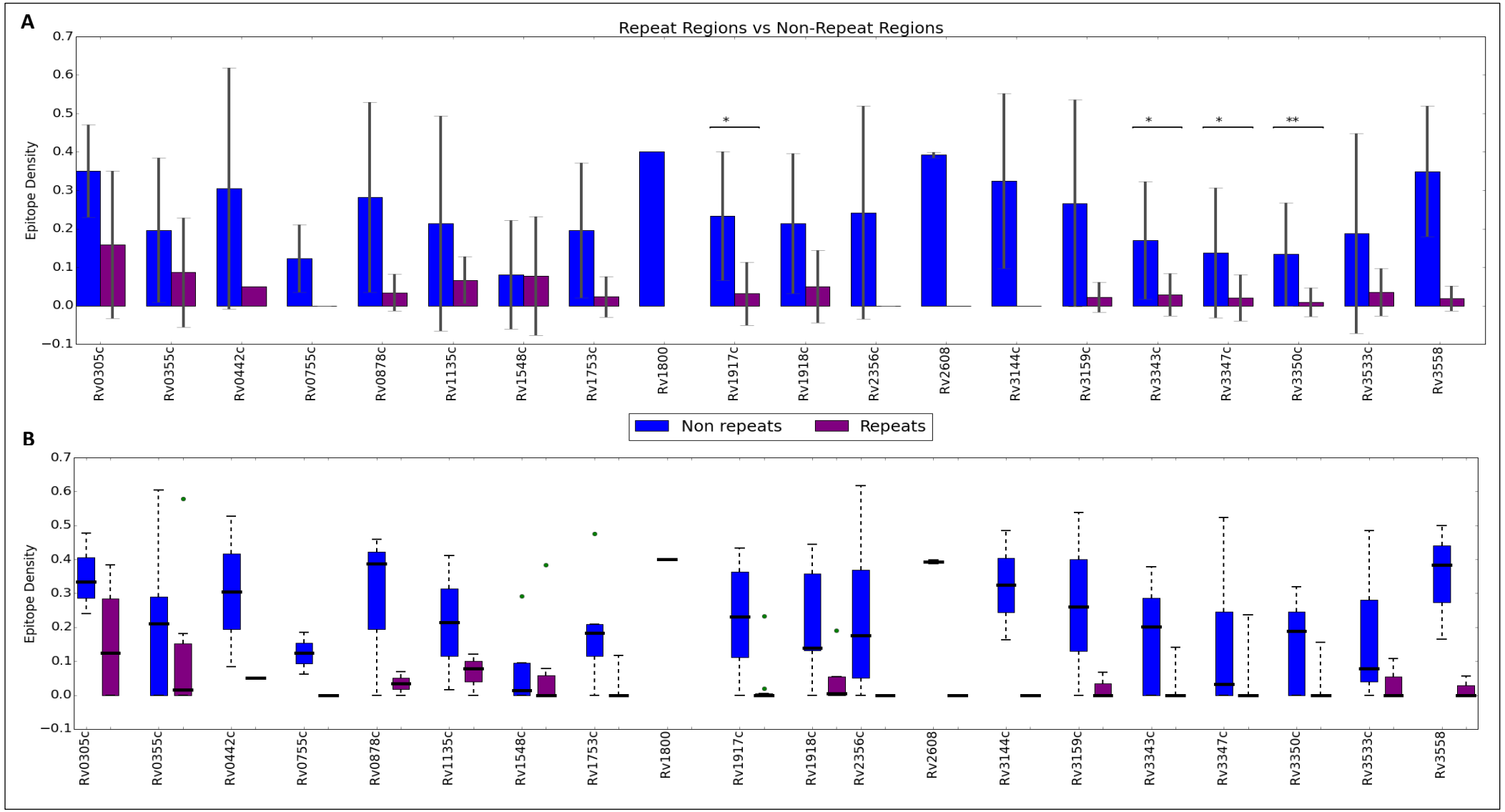


Figure 5.6: Comparison of epitope density within repeat and non-repeat regions individually for each PPE_MPTR protein. A) Mean epitope density within non-repeated regions (blue) is compared to repeat regions (purple) for each PPE_MPTR protein. Error bars show standard deviation in epitope density. Significant differences are shown where applicable (* $p < 0.05$, ** $p < 0.01$). B) Box plot showing the distribution of epitope density in non-repeat regions (blue) versus repeat regions (purple), with outliers shown in green.

Mean epitope density is higher in non-repeat regions for all PPE_MPTR proteins. Given that certain proteins have only one or two repeat regions along the length of the protein, the sample size for a statistical test is too small and therefore even though it is evident that epitope density is higher in repeat versus non-repeat regions, statistical significance cannot be reported. Rv1800 has no repeat regions and one of the highest overall epitope densities across all PPE_MPTR proteins. Certain proteins have a zero epitope density in repeat regions (Rv0755c, Rv2356c, Rv2608, and Rv3144c). Two proteins have outliers within the repeat regions, Rv0305c and Rv1548c, indicating that there are specific repeat regions within these two proteins that have a high epitope density.

5.4 Conclusion

This chapter has explored the genetic variation of the PPE_MPTR proteins and how this potentially affects epitope density and therefore interaction with the host immune system. The types of genetic variation identified include IS*6110* insertions, a variable copy number of tandem repeats and micro-mutations such as SNP's and small indels. When comparing results for individual PPE_MPTR proteins, it is clear that not all *ppe_mptr* genes are equally variable as previously suggested in literature (Chapter 1), but rather that certain members are highly variable (Rv0355c, Rv1753c and Rv1917c), while others are relatively conserved across strains with either no or at most one micro-mutation identified, no IS*6110* insertions and a constant number of MPTR repeats (Rv0442c, Rv0755c, Rv0878c, Rv1548c, Rv2356c, Rv2608c, and Rv3533c). This may indicate different functional roles and/or importance for the PPE_MPTR proteins, and a possible further sub division of the PPE_MPTR family based on either the type of variation (e.g. those with longer non-MPTR tandem repeats) or based on potential functional roles. The reason and potential functional impact of a different number of tandem repeats is still currently unknown and should be further explored.

Other genetic variation results from this chapter that should be further explored include:

- The role of IS*6110* insertion elements within the *ppe_mptr* genes and the impact on protein structure and function as a result of the insertions.

- The relationship between Rv3343c, Rv3350c and Rv3347c and their presence within different *M. tuberculosis* strains. There is a high sequence similarity between these three genes, making it difficult to identify which ortholog(s) are present in specific strains. These genes may represent gene duplication events and/or recombination hotspots.

- The importance of the highly variable protein Rv1917c. This is by far the most variable of all the PPE_MPTR proteins, containing IS*6110* insertions elements, and a variable copy number of both MPTR and longer non-MPTR repeats. Possible partial gene deletions which make multiple alignments difficult may also be present. It is interesting to compare Rv1917c and Rv1918c, which are close in proximity within the genome to each other, and both contain MPTR and longer non-MPTR repeats. However Rv1917c is highly variable while Rv1918c is conserved across strains. The relationship between these proteins and whether one arose as a duplication event of the other and therefore whether they serve similar roles within the pathogenesis of the bacteria should be investigated. This may provide a reason for a lack of selective pressure on Rv1917c and the resulting high variability evident. Rv1917c has been shown to be cell wall associated and surface exposed (Sampson *et al*. 2001), and is involved in inducing host cell signalling and promoting Th2 response (Bansal *et al*. 2010), which may be evidence that Rv1917c may indeed differently modulate host immune response.

The effect of micro-mutations such as SNP's and indels on epitope density has been investigated. However a clear pattern of micro-mutations affecting HLA class II binding ability could not be identified. Investigation of repeat and non-repeat regions has shown a correlation with the fluctuating patterns of epitopes along the length of the protein, with high epitope density seen in non-repeat regions. Overall these findings suggest that epitopes within the PPE_MPTR proteins is (in general) restricted to conserved non-repeating regions. No evidence to support antigenic variation or that genetic diversity is modulating host-pathogen interactions

was found. Previous literature however has shown that the PPE_MPTR proteins are differentially expressed in *M. tuberculosis in vivo* (Soldini *et al*. 2011), and that the gene expression pattern in different types of tissue (lung and spleen) is different for each *ppe_mptr* gene. The *ppe_mptr* genes may therefore respond differently depending on the complex environmental signals within various host tissues (Soldini *et al*. 2011).

Understanding the genetic variation of the *ppe_mptr* genes is an extremely important component of deciphering the role of these proteins within *M. tuberculosis* pathogenesis, which is currently largely unknown. Given the technical difficulties in sequencing and aligning these genes, previous studies have often excluded them in other genetic variation analyses. Therefore, while the technical complications may still be a limitation within this investigation that should be kept in mind, to date it is the most comprehensive analysis of PPE_MPTR variation. In addition, the small number of strains within this study may also be a limitation, and the genetic variation therefore underestimated. A larger number of strains, and high quality next generation sequencing data could provide additional information not reported here. Targeted Sanger sequencing of specific variations could also be used to validate the genetic variation identified. Further studies could also take into consideration possible differences in the expression profile of PPE_MPTR proteins with different copy numbers of tandem repeats in various biological conditions similarly to the study performed by Soldini *et al*. (2011). The novel approaches used within this chapter to overcome the technical difficulties associated with mapping and aligning the *ppe_mptr* genes could also be applied to other variable and/or repeat regions of the *M. tuberculosis* genome such as the *pe_pgrs* genes.

Comparative genomics is an essential part of the reverse vaccinology pipeline, especially for *M. tuberculosis* (Chapter 2). The results from this chapter will therefore be important in identifying potential vaccine candidates. Ideally vaccine candidates should be conserved across strains in order to be affective in diverse populations affected with strains from various lineages. The more conserved PPE_MPTR proteins may therefore be superior to the highly variable PPE_MPTR proteins. This is further explored in Chapter 6.

5.5 References

Bansal, K. *et al*., 2010. Src homology 3-interacting domain of Rv1917c of *Mycobacterium tuberculosis* induces selective maturation of human dendritic cells by regulating PI3K-MAPK-NF-kappaB signaling and drives Th2 immune responses. *The Journal of biological chemistry*, 285(47), pp.36511–22.

Benson, G., 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic acids research*, 27(2), pp.573–80.

Cock, P.J.A. *et al*., 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11), pp.1422–1423.

Cole, S.T. *et al*., 1998. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature*, 393(6685), pp.537–44.

Copin, R. *et al*., 2014. Sequence diversity in the pe_pgrs genes of *Mycobacterium tuberculosis* is independent of human T cell recognition. *mBio*, 5(1), pp.e00960–13.

Darling, A.C.E. *et al*., 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome research*, 14(7), pp.1394–403.

Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, 32(5), pp.1792–7.

Galagan, J.E. *et al*., 2010. TB database 2010: overview and update. *Tuberculosis (Edinburgh, Scotland)*, 90(4), pp.225–35.

Iwai, H. *et al*., 2015. CASTB (the comprehensive analysis server for the *Mycobacterium tuberculosis* complex): A publicly accessible web server for epidemiological analyses, drug-resistance prediction and phylogenetic comparison of clinical isolates. *Tuberculosis (Edinburgh, Scotland)*, 95(6), pp.843–4.

Lew, J.M. *et al*., 2011. TubercuList--10 years after. *Tuberculosis (Edinburgh, Scotland)*, 91(1), pp.1–7.

Liu, X. *et al*., 2006. Evidence for recombination in *Mycobacterium tuberculosis*. *Journal of bacteriology*, 188(23), pp.8169–77.

McEvoy, C.R.E. *et al*., 2012. Comparative analysis of *Mycobacterium tuberculosis* pe and ppe genes reveals high sequence variation and an apparent absence of selective constraints. *PLoS ONE*, 7(4).

McEvoy, C.R.E. *et al*., 2009. Evidence for a rapid rate of molecular evolution at the hypervariable and immunogenic *Mycobacterium tuberculosis* PPE38 gene region. *BMC evolutionary biology*, 9, p.237.

Needleman, S.B. & Wunsch, C.D., 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3), pp.443–53.

Nei, M., 1987. *Molecular evolutionary genetics.*, New York, NY: Columbia University Press.

Sampson, S.L. *et al*., 2001. Expression, characterization and subcellular localization of the *Mycobacterium tuberculosis* PPE gene Rv1917c. *Tuberculosis (Edinburgh, Scotland)*, 81(5-6), pp.305–17.

Soldini, S. *et al*., 2011. PPE-MPTR genes are differentially expressed by *Mycobacterium tuberculosis* in vivo. *Tuberculosis*, 91(6), pp.563–568.

Tamura, K. *et al*., 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular biology and evolution*, 28(10), pp.2731–9.

Tatusova, T. *et al*., 2016. NCBI prokaryotic genome annotation pipeline. *Nucleic acids research*, 44(14),

pp.6614–24.

Teixeira-Silva, A. *et al*., 2013. The role of recombination in the origin and evolution of Alu subfamilies. *PloS one*, 8(6), p.e64884.

Treangen, T.J. & Salzberg, S.L., 2012. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature reviews. Genetics*, 13(1), pp.36–46.

Chapter 6: Identification of Potential Vaccine Candidates

6.1 Introduction

There is a crucial need for a new vaccine against *M. tuberculosis,* and reverse vaccinology (RV) approaches have been used extensively to search for new sub-unit vaccine candidates within the *M. tuberculosis* proteome (Chapter 2). Different selection criteria (i.e. by type or function of protein) are often used to first narrow down the list of possible antigenic targets being fed into a RV workflow. This thesis has focused on a particular family of proteins, the PPE_MPTR proteins. PE/PPE proteins have previously been used as a starting point within RV workflows, and even when whole genome approaches are used, the PE/PPE proteins are often in the list of resulting potential vaccine candidates (Chapter 2). Certain members of the PE/PPE protein family have also already been included in current vaccines in clinical trials (Chapter 1, Table 1.1). However, to date, no studies have focused specifically on the PPE_MPTR subfamily, even though they have been hypothesized to play a role in host-pathogen interactions.

Reverse vaccinology workflows include comparative genomics, epitope prediction, analysis of protein properties and the prediction of potential population coverage (Chapter 2, Figure 2.2). CD4+ T-cell epitope prediction has been performed for the PPE_MPTR proteins (Chapter 3) and a large number of potential epitopes have been identified, supporting the hypothesis of a potential role for the PPE_MPTR proteins within host-pathogen interactions. The PPE_MPTR proteins have also been speculated to play a role in antigenic variation, allowing the bacteria to evade the host immune system. The hypothesis of whether or not genetic variation within the PPE_MPTR proteins differentially modulates host immune response and therefore supporting the speculation of antigenic variation has been explored in Chapter 5. The analysis of genetic diversity within the PPE_MPTR proteins across various *M. tuberculosis* strains shows a correlation between areas of repeat regions and epitope density but no evidence for the PPE_MPTR proteins playing a role in antigenic variation was found. The results however do show that certain members of the PPE_MPTR proteins are highly variable while others are more conserved. Comparative genomics is an essential step within RV workflows, as a potential vaccine candidate should be conserved across strains in order to be affective in diverse populations. This chapter uses the results from epitope prediction (Chapter 4) and genetic diversity (Chapter 5) to filter potential vaccine candidates within the PPE_MPTR proteins. In addition, the potential population coverage within high TB burden countries in Africa of possible vaccine candidates has been investigated in order to further narrow down the list of potential candidates.

6.2 Methods

Figure 6.1 shows the filtering methodology used within a customised RV workflow in order to identify potential vaccine candidates within the PPE_MPTR protein family.

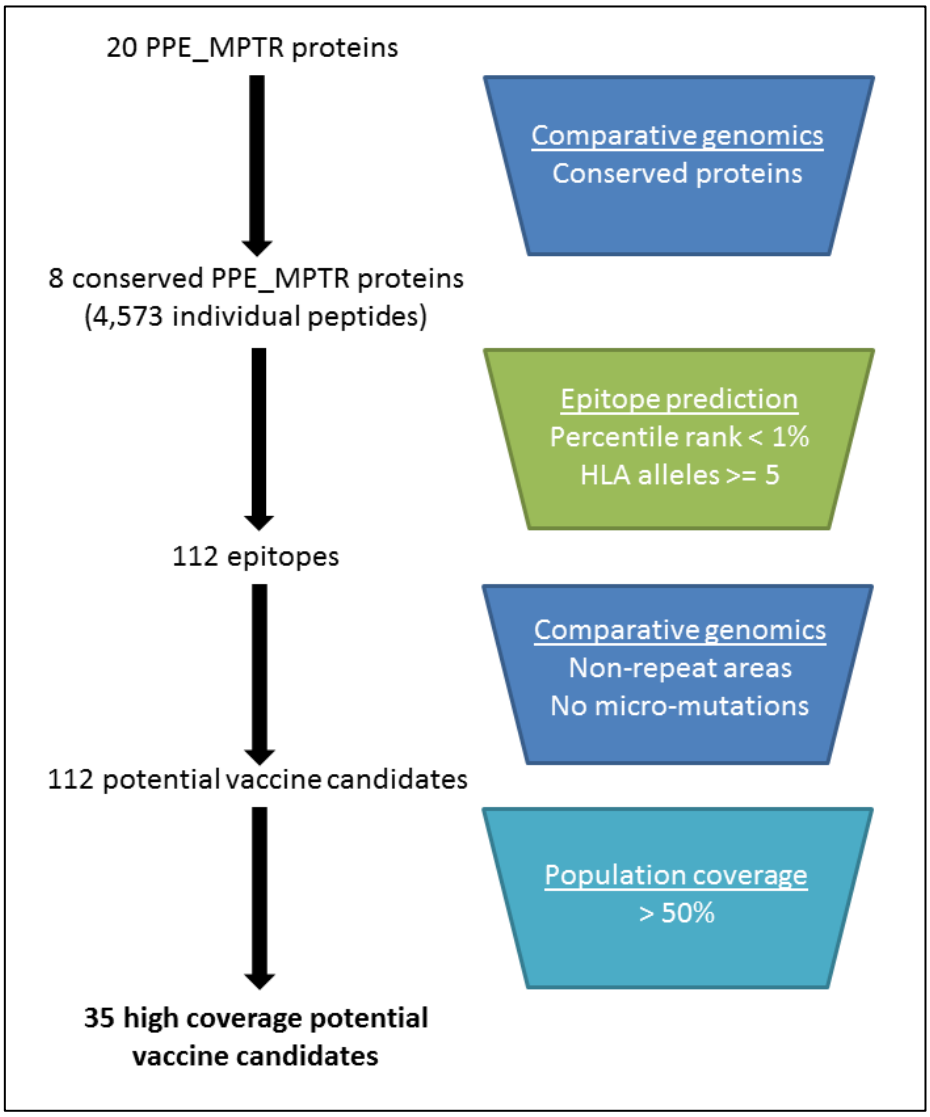


Figure 6.1: RV workflow used for the PPE_MPTR proteins. Comparative genomics results from Chapter 5 and epitope prediction results from Chapter 4 are used to filter the overlapping 15mer peptides within the PPE_MPTR proteins. Potential population coverage is calculated to identify high coverage vaccine candidates within the high TB burden countries in Africa.

Similarly to Chapter 4 and Chapter 5, three of the PPE_MPTR proteins were not included in the analysis due to dissimilarity in protein sequence to the remaining PPE_MPTR proteins. The starting point for the RV workflow therefore included 20 PPE_MPTR proteins. Results from Chapter 5 were used to filter these proteins for those that are highly conserved across strains. Highly conserved proteins are defined as those with either no or at most one micro-mutation identified, no IS*6110* insertion elements, a constant number of MPTR repeats and no long non-MPTR repeats (Rv0442c, Rv0755c, Rv0878c, Rv1548c, Rv2356c, Rv2608c, and Rv3533c). Rv1800 has also been included even though more than 1 micro-mutation was identified due to the high epitope density within this protein (Chapter 4), and no repeat regions (Chapter 5). Epitope prediction results from Chapter 4 were used to identify peptides within these proteins that are predicted promiscuous epitopes (percentile rank < 1%, >= 5 HLA class II alleles). Results from Chapter 5 revealed that epitope dense regions correlated with non-repeat areas, however there were certain outliers where epitopes were found within repeat regions (Rv1548c). An extra filtering step has been included to remove possible vaccine candidates within repeat regions, as well as any epitope affected by a micro-mutation.

An epitope clustering analysis was performed on the filtered epitopes using the epitope cluster analysis tool available within the IEDB analysis resource (Kim et al. 2012), which groups epitopes into clusters based on their sequence identity. Different identity thresholds were used to determine clusters (70% - 100%).

Potential population coverage for the filtered epitopes was calculated using the IEDB population coverage tool (Kim et al. 2012; Bui et al. 2006). Population coverage was calculated for 9 African countries within the 22 high burden TB countries for which HLA class II frequency information is available. The IEDB population coverage tool uses HLA class II frequencies from the allelefrequencies.net database (González-Galarza et al. 2015). Epitopes that have greater than 50% population coverage in at least one country were identified as possible vaccine candidates.

6.3 Results

6.3.1 Results summary

Table 6.1 shows a summary of the resulting number of potential vaccine candidates after each filtering step.

Table 6.1: Summary results of potential vaccine candidates. The number of potential epitopes after each filtering step is shown for the 8 conserved PPE_MPTR proteins used in the RV pipeline.

Protein	Number of peptides	Predicted epitopes (Percentile rank > 1%)	Promiscuous Epitopes (HLA alleles >= 5)		Non-repeat conserved epitopes*	>50% population coverage in at least one African high burden TB country
			PPE	MPTR		
Rv0442c	473	115	5	-	5	-
Rv0755c	631	69	7	-	7	4
Rv0878c	429	105	-	7	7	7
Rv1548c	664	76	3	-	3	-
Rv1800	641	257	-	43	43	11
Rv2356c	601	105	8	-	8	2
Rv2608	566	195	-	25	25	8
Rv3533c	568	108	12	2**	14	3
Total	**4,573**	**1030**	**112**		**112**	**35**

*None of the predicted promiscuous epitopes where found in repeat regions or were affected by micro-mutations.
** Overlapping PPE and MPTR boundary.

Of the 4,573 overlapping 15mer peptides found within the 8 PPE_MPTR proteins included in the RV workflow, 1,030 are predicted epitopes able to bind to at least one of the 47 HLA class II alleles included in the epitope prediction (Chapter 4). Of these, 112 are able to bind to at least 5 HLA class II alleles. None of these 112 promiscuous epitopes were found within repeated regions, nor had micro-mutations affecting the amino acid sequence in these regions, further emphasizing that *M. tuberculosis* epitopes are conserved. After further filtering for population coverage in high burden TB countries in Africa, 35 unique high coverage vaccine candidates remain.

6.3.2 Epitope cluster analysis

An epitope cluster analysis was performed using the 112 promiscuous epitopes identified before filtering for population coverage, in order to determine the uniqueness of the collection of epitopes. A high similarity between epitopes would collapse the number of potential vaccine candidates. Different sequence similarity thresholds were investigated when determining clustering.

- When using a threshold of 90% and 100%, no clusters were found, indicating that no two epitopes within the 112 epitopes identified were the exactly the same, or even 90% similar.

- When using a threshold of 80%, one cluster containing only two sequences was identified, indicating that two out of the 112 sequences are 80% similar, and the remaining 110 epitopes are more than 20% dissimilar to each other. These two epitopes come from Rv0755c and Rv3533c.
- When using a threshold of 70%, 7 clusters each containing only 2 sequences where identified. These are shown in Table 6.2 below.

Table 6.2: Epitope cluster analysis at 70% similarity threshold. Seven clusters each containing two epitopes were identified to have 70% similarity.

Cluster	Protein	Position	Epitope
1	Rv3533c	9	ELNSLRMFTGAGSAP
	Rv0755c	10	ETNSLRMYLGAGSRP
2	Rv3533c	11	NSLRMFTGAGSAPML
	Rv0755c	12	NSLRMYLGAGSRPLL
3	Rv0755c	13	SLRMYLGAGSRPLLA
	Rv3533c	12	SLRMFTGAGSAPMLA
4	Rv3533c	13	LRMFTGAGSAPMLAA
	Rv0755c	14	LRMYLGAGSRPLLAA
5	Rv0755c	8	PPETNSLRMYLGAGS
	Rv3533c	7	PPELNSLRMFTGAGS
6	Rv0755c	9	PETNSLRMYLGAGSR
	Rv3533c	8	PELNSLRMFTGAGSA
7	Rv3533c	107	PLLVAANRNAFAQLV
	Rv2356c	111	PVVVAANRSAFVQLV

Epitopes from Rv0755c and Rv3533c clustered together in six out of the seven clusters. These epitopes are in similar positions within the respective proteins indicating a similarity between Rv0755c and Rv3533c within this region. One cluster contained an epitope from Rv3533c and one from Rv2356c, also at similar positions within the respective proteins.

Given the small number of clusters identified, these results show the uniqueness of the 112 epitopes identified within the PPE_MPTR proteins and that the majority of the epitopes are more than 30% dissimilar to one another.

6.3.3 Potential population coverage

Potential population coverage was calculated for each of the 112 promiscuous epitopes for high TB burden countries in Africa. These countries include Zimbabwe, Uganda, Tanzania, South Africa, Nigeria, Mozambique, Kenya, Ethiopia and Congo. HLA class II frequencies from the allelefrequencies.net database are used to calculate population coverage. The information within the allelefrequencies.net database is sourced from various literature and relies on contributions made by researchers. Where studies on the frequency of HLA alleles in specific countries have not been previously performed, no information for these countries will be available within the database and population coverage cannot be calculated. Information within the database is therefore not fully inclusive of all countries. Where information is available for a specific country, it should be kept in mind that this information may not be fully comprehensive for all of the possible different ethnic groups living within that country, as data may originate from studies specifically focused on one geographical area or specific ethnicity that may not be fully representative of the country as a whole. Where more than one study within a country is available, or different ethnic groups have been included, these different ethnic populations are indicated within the database and population coverage can be calculated separately. The HLA allele frequencies for Tanzania and Mozambique are not available within the allelefrequencies.net database and these countries were therefore not included within this analysis.

Potential population coverage for all 112 epitopes was 0% for Uganda, Nigeria, and Kenya. This may be due to limited HLA allele frequency data available within the allelefrequencies.net database for these countries. Results for Zimbabwe, South Africa, Ethiopia and the Congo are shown in Appendix A. The list of 112

epitopes was further filtered to include only those epitopes that had greater than 50% coverage in at least one of the four African countries for which results are available. This narrowed down the list to 35 high coverage potential vaccine candidates (Table 6.3).

Table 6.3: High coverage potential vaccine candidates in high TB burden countries in Africa. The predicted population coverage for each epitope within each of the four African countries, Zimbabwe, South Africa, Ethiopia and Congo is shown. Coverage > 50% is shown is highlighted in red.

Protein	Position	Epitope	Zimbabwe	South Africa (Black)	Ethiopia	Congo
Rv0755c	11	TNSLRMYLGAGSRPL	48.73%	23.96%	62.88%	47.71%
Rv0755c	12	NSLRMYLGAGSRPLL	48.73%	23.96%	62.88%	47.71%
Rv0755c	13	SLRMYLGAGSRPLLA	53.90%	24.66%	66.19%	51.61%
Rv0755c	14	LRMYLGAGSRPLLAA	53.90%	24.66%	66.19%	51.61%
Rv0878c	415	TARECSIRVIISRVS	62.67%	23.96%	36.48%	53.31%
Rv0878c	416	ARECSIRVIISRVSS	63.52%	23.96%	36.48%	53.31%
Rv0878c	417	RECSIRVIISRVSST	72.44%	45.98%	53.90%	64.41%
Rv0878c	418	ECSIRVIISRVSSTG	71.48%	45.98%	50.11%	60.84%
Rv0878c	419	CSIRVIISRVSSTGA	72.44%	45.98%	53.90%	64.41%
Rv0878c	420	SIRVIISRVSSTGAP	70.73%	45.98%	50.11%	60.84%
Rv0878c	421	IRVIISRVSSTGAPP	70.73%	45.98%	50.11%	60.84%
Rv1800	425	TMTQYYIIRTENLPL	29.10%	5.91%	77.58%	44.29%
Rv1800	426	MTQYYIIRTENLPLL	35.20%	6.68%	80.14%	48.32%
Rv1800	427	TQYYIIRTENLPLLE	35.20%	6.68%	80.14%	48.32%
Rv1800	428	QYYIIRTENLPLLEP	29.10%	5.91%	76.00%	42.63%
Rv1800	491	PEVSPVVIADALVAG	53.08%	45.98%	46.91%	49.76%
Rv1800	492	EVSPVVIADALVAGT	53.08%	45.98%	46.91%	49.76%
Rv1800	493	VSPVVIADALVAGTQ	53.08%	45.98%	46.91%	49.76%
Rv1800	494	SPVVIADALVAGTQQ	53.08%	45.98%	46.91%	49.76%
Rv1800	632	GGLQLLIIISAGRTI	53.62%	0.00%	36.50%	42.16%
Rv1800	633	GLQLLIIISAGRTIA	72.12%	0.00%	50.18%	42.16%
Rv1800	634	LQLLIIISAGRTIAN	72.12%	0.00%	50.18%	42.16%
Rv2356c	118	RSAFVQLVLSNVFGQ	57.75%	23.96%	29.94%	50.04%
Rv2356c	119	SAFVQLVLSNVFGQN	57.75%	23.96%	29.94%	50.04%
Rv2608	346	GNEVVVFGTSQSATI	40.71%	23.96%	50.76%	45.89%
Rv2608	359	TIATFEMRYLQSLPA	62.67%	23.96%	42.85%	53.31%
Rv2608	360	IATFEMRYLQSLPAH	67.05%	24.66%	46.97%	56.99%
Rv2608	361	ATFEMRYLQSLPAHL	91.99%	26.21%	90.95%	77.47%
Rv2608	362	TFEMRYLQSLPAHLR	95.29%	40.86%	90.95%	80.62%
Rv2608	363	FEMRYLQSLPAHLRP	95.29%	40.86%	90.95%	80.62%
Rv2608	364	EMRYLQSLPAHLRPG	94.29%	40.86%	86.33%	78.64%
Rv2608	365	MRYLQSLPAHLRPGL	78.47%	40.86%	46.97%	67.48%
Rv3533c	106	HPLLVAANRNAFAQL	40.71%	58.40%	26.01%	32.83%
Rv3533c	7	PPELNSLRMFTGAGS	64.48%	25.52%	36.48%	56.40%
Rv3533c	8	PELNSLRMFTGAGSA	64.48%	25.52%	36.48%	56.40%

The population coverage of the 35 vaccine candidates within these four African countries varies substantially when comparing one country to another, indicating that the best combination of epitopes to be included in a vaccine cocktail could be different for different countries. Alternatively a vaccine cocktail should include a range of epitopes that provide high coverage within different populations. This would be further exemplified if all 22 high burden TB countries were included within a population coverage analysis. No potential epitope had population coverage of at least 50% in all four of the African countries investigated, with only one candidate (within Rv3533c) with population coverage of more than 50% in South Africa.

6.4 Conclusion

This chapter has used results from epitope prediction (Chapter 4), comparative genomics (Chapter 5), and potential population coverage to identify possible vaccine candidates within the PPE_MPTR proteins. Thirty five high coverage vaccine candidates within high burden TB countries within Africa have been identified. When comparing the population coverage between Zimbabwe, South Africa, Ethiopia and Congo, a significant difference can be seen. This is due to the difference in frequencies between HLA alleles in the various populations. This result reiterates the vital importance of using various populations to identify vaccine targets as well as using various populations within clinical trials, as a vaccine that shows promising results in one specific population may not be effective when used in another population in a distinct geographical location. Even within one continent such as Africa, HLA allele frequencies can differ substantially between countries, which will affect the potential efficacy of a vaccine within those countries. This analysis has also shown the importance of having high quality HLA frequency data in all countries, and within different ethnic groups within those countries. Currently the allefrequencies.net database is not fully comprehensive of all countries and ethnicities. Using HLA allele information, host-customised vaccines may offer a solution to increasing a vaccine's effectiveness. Alternatively, a vaccine cocktail can contain epitopes from various proteins that have high coverage within different populations in order to maximise overall (global) population coverage.

Only one epitope (out of the 112) had predicted population coverage of greater than 50% in South Africa. Vaccine candidates containing epitopes from the PPE_MPTR proteins may therefore not be as effective in South Africa compared to other countries, or should be supplemented with epitopes from other proteins in order to increase coverage within South Africa. A large proportion of the potential vaccine candidates identified had greater than 50% coverage in the other three African countries, with certain epitopes having over 90% coverage. The cut-off of 50% used when filtering epitopes can be increased to further narrow down the potential list of epitopes to be validated in a wet lab setting. The sensitivity and specificity of the population coverage tools has not been extensively investigated within the literature, as validated data about an epitope's population coverage is scarcely available. The exact cut-off percentage to use therefore varies between individual studies.

The results presented within this chapter support the hypothesis presented in Chapter 1 (Section 1.3) that epitopes within the PPE_MPTR proteins may be possible subunit vaccine candidates for *M. tuberculosis.* Further studies should also include possible CD8+ T-cell epitopes. These results are based on *in silico* techniques obtained using a reverse vaccinology approach and the thirty five vaccine candidates should be fed into an experimental wet lab validation. Experimental validation can include using binding assays to test for HLA class II binding, proliferation of CD4+ T-cells, or testing for IFN-γ, IL-2 and TNF-α response.

Results from this thesis reiterate the important role a computational approach can play when trying to determine the function of a currently unknown protein or family of proteins. In this case, the possible role of the PPE_MPTR proteins within host-pathogen interactions was investigated and whether or not possible sub-unit vaccine candidates exist within the PPE_MPTR proteins explored.

The main conclusions based on the findings from this thesis include:

- CD4+ T-cell epitopes exist within the PPE_MPTR proteins, supporting the hypothesis of host-pathogen interactions for the PPE_MPTR proteins.
- Certain members of the PPE_MPTR proteins are genetically variable while others are more conserved across strains, possibly indicating different functional roles.
- Genetically diverse regions are less likely to contain epitopes, and no evidence to support antigenic variation was found.
- Areas of high and low epitope density are correlated to areas of non-repeat and repeat regions within the genome respectively. Epitopes within the PPE_MPTR proteins are therefore conserved non-repeating peptides.
- There are possible vaccine candidates with high predicted population coverage in African countries within the PPE_MPTR proteins.

6.5 References

Bui, H. et al., 2006. Predicting population coverage of T-cell epitope-based diagnostics and vaccines. *BMC bioinformatics*, 7, p.153.

González-Galarza, F.F. et al., 2015. Allele frequency net 2015 update: new features for HLA epitopes, KIR and disease and HLA adverse drug reaction associations. *Nucleic acids research*, 43(Database issue), pp.D784–8.

Kim, Y. et al., 2012. Immune epitope database analysis resource. *Nucleic acids research*, 40(Web Server issue), pp.W525–30.

6.6. Appendix

A: Population coverage for high burden TB countries in Africa

For each of the 112 promiscuous epitopes identified, the source protein and position within the protein is shown, the HLA alleles the epitope is predicted to bind to, and the predicted population coverage within 4 high burden TB countries in Africa is shown.

Table 6.4: Population coverage of potential vaccine candidates in high TB burden countries in Africa. HLA alleles predicted to bind to each epitope (Results from chapter 4), are used to calculate the potential population coverage within four African countries with high TB burden.

Protein	Pos	Epitope	HLA Alleles	Zim-babwe	South Africa (Black)	Ethiopia	Congo
Rv0442c	154	MAGYHFDASAAVAQL	*HLA-DQA1*04:01/DQB1*04:02, HLA-DRB1*09:01, HLA-DRB3*01:01, HLA-DQA1*03:01/DQB1*03:02, HLA-DRB1*04:01*	2.98%	1.79%	3.25%	6.68%
Rv0442c	155	AGYHFDASAAVAQLA	*HLA-DQA1*04:01/DQB1*04:02, HLA-DRB1*09:01, HLA-DRB3*01:01, HLA-DQA1*03:01/DQB1*03:02, HLA-DQA1*01:02/DQB1*06:02, HLA-DRB1*04:01*	2.98%	1.79%	3.25%	6.68%
Rv0442c	72	ATQYLAWLSTAAAQA	*HLA-DQA1*04:01/DQB1*04:02, HLA-DRB1*09:01, HLA-DRB1*08:02, HLA-DQA1*03:01/DQB1*03:02, HLA-DRB1*10:01*	10.13%	2.58%	5.48%	9.84%
Rv0442c	73	TQYLAWLSTAAAQAE	*HLA-DQA1*04:01/DQB1*04:02, HLA-DRB1*09:01, HLA-DRB1*08:02, HLA-DQA1*03:01/DQB1*03:02, HLA-DRB1*10:01*	10.13%	2.58%	5.48%	9.84%
Rv0442c	97	ATAFEAALAATVQPA	*HLA-DQA1*04:01/DQB1*04:02, HLA-DRB1*09:01, HLA-DQA1*05:01/DQB1*03:01, HLA-DQA1*03:01/DQB1*03:02, HLA-DRB1*01:01*	6.88%	1.79%	0.00%	8.80%
Rv0755c	10	ETNSLRMYLGAGSRP	*HLA-DRB1*11:02, HLA-DRB1*01:02, HLA-DRB1*15:02, HLA-DRB1*13:01, HLA-DRB1*08:04*	41.94%	23.96%	32.87%	37.86%
Rv0755c	11	TNSLRMYLGAGSRPL	*HLA-DRB1*11:02, HLA-DRB1*07:01, HLA-DRB1*01:02, HLA-DRB1*15:02, HLA-DRB1*13:01, HLA-DRB1*08:04*	48.73%	23.96%	62.88%	47.71%
Rv0755c	12	NSLRMYLGAGSRPLL	*HLA-DRB1*11:02, HLA-DRB1*07:01, HLA-DRB1*01:02, HLA-DRB1*15:02, HLA-DRB1*13:01, HLA-DRB1*08:04*	48.73%	23.96%	62.88%	47.71%
Rv0755c	13	SLRMYLGAGSRPLLA	*HLA-DRB1*11:02, HLA-DRB1*07:01, HLA-DRB1*10:01, HLA-DRB5*01:01, HLA-DRB1*01:02, HLA-DRB1*15:02, HLA-DRB1*13:01, HLA-DRB1*08:04*	53.90%	24.66%	66.19%	51.61%
Rv0755c	14	LRMYLGAGSRPLLAA	*HLA-DRB1*11:02, HLA-DRB1*07:01, HLA-DRB1*10:01, HLA-DRB5*01:01, HLA-DRB1*01:02, HLA-DRB1*15:02, HLA-DRB1*13:01, HLA-DRB1*08:04*	53.90%	24.66%	66.19%	51.61%
Rv0755c	8	PPETNSLRMYLGAGS	*HLA-DRB1*11:02, HLA-DRB1*01:02, HLA-DRB1*15:02, HLA-DRB1*13:01, HLA-DRB1*08:04*	41.94%	23.96%	32.87%	37.86%
Rv0755c	9	PETNSLRMYLGAGSR	*HLA-DRB1*11:02, HLA-DRB1*01:02, HLA-DRB1*15:02, HLA-DRB1*13:01, HLA-DRB1*08:04*	41.94%	23.96%	32.87%	37.86%
Rv0878c	415	TARECSIRVIISRVS	*HLA-DRB5*01:01, HLA-DRB1*01:02, HLA-DRB1*08:04, HLA-DRB1*11:04, HLA-DRB1*11:02, HLA-DRB1*13:01, HLA-DRB1*11:01*	62.67%	23.96%	36.48%	53.31%

Protein	Pos	Epitope	HLA Alleles	Zim-babwe	South Africa (Black)	Ethiopia	Congo
Rv0878c	416	ARECSIRVIISRVSS	*HLA-DRB5*01:01, HLA-DRB1*01:02, HLA-DRB1*12:02, HLA-DRB1*08:04, HLA-DRB1*11:04, HLA-DRB1*11:02, HLA-DRB1*13:01, HLA-DRB1*11:01*	63.52%	23.96%	36.48%	53.31%
Rv0878c	417	RECSIRVIISRVSST	*HLA-DRB5*01:01, HLA-DRB1*01:02, HLA-DRB1*13:03, HLA-DRB1*12:02, HLA-DRB1*08:04, HLA-DRB1*11:04, HLA-DRB1*11:02, HLA-DRB1*08:03, HLA-DRB1*11:03, HLA-DRB1*13:01, HLA-DRB1*03:01, HLA-DRB1*11:01*	72.44%	45.98%	53.90%	64.41%
Rv0878c	418	ECSIRVIISRVSSTG	*HLA-DRB5*01:01, HLA-DRB1*01:02, HLA-DRB1*12:02, HLA-DRB1*08:04, HLA-DRB1*11:04, HLA-DRB1*11:02, HLA-DRB1*08:03, HLA-DRB1*11:03, HLA-DRB1*13:01, HLA-DRB1*08:02, HLA-DRB1*03:01, HLA-DRB1*11:01*	71.48%	45.98%	50.11%	60.84%
Rv0878c	419	CSIRVIISRVSSTGA	*HLA-DRB5*01:01, HLA-DRB1*01:02, HLA-DRB1*13:03, HLA-DRB1*12:02, HLA-DRB1*08:04, HLA-DRB1*11:04, HLA-DRB1*11:02, HLA-DRB1*08:03, HLA-DRB1*11:03, HLA-DRB1*13:01, HLA-DRB1*08:02, HLA-DRB1*03:01, HLA-DRB1*14:04, HLA-DRB1*11:01*	72.44%	45.98%	53.90%	64.41%
Rv0878c	420	SIRVIISRVSSTGAP	*HLA-DRB1*01:02, HLA-DRB1*08:04, HLA-DRB1*11:04, HLA-DRB1*11:02, HLA-DRB1*08:03, HLA-DRB1*13:01, HLA-DRB1*08:02, HLA-DRB1*03:01, HLA-DRB1*11:01*	70.73%	45.98%	50.11%	60.84%
Rv0878c	421	IRVIISRVSSTGAPP	*HLA-DRB1*01:02, HLA-DRB1*08:04, HLA-DRB1*11:04, HLA-DRB1*11:02, HLA-DRB1*13:01, HLA-DRB1*08:02, HLA-DRB1*03:01, HLA-DRB1*11:01*	70.73%	45.98%	50.11%	60.84%
Rv1548c	112	ANRGRLRSLVASNLL	*HLA-DRB1*04:03, HLA-DRB1*04:06, HLA-DRB1*01:01, HLA-DRB1*10:01, HLA-DRB1*01:02*	22.21%	0.80%	29.64%	16.09%
Rv1548c	113	NRGRLRSLVASNLLG	*HLA-DRB1*04:03, HLA-DRB1*04:06, HLA-DRB1*01:01, HLA-DRB1*10:01, HLA-DRB1*01:02*	22.21%	0.80%	29.64%	16.09%
Rv1548c	114	RGRLRSLVASNLLGQ	*HLA-DRB1*04:03, HLA-DRB1*04:06, HLA-DRB1*11:04, HLA-DRB1*01:01, HLA-DRB1*10:01, HLA-DRB1*01:02*	23.44%	0.80%	29.64%	20.08%
Rv1800	168	ASWLQRLQSIPGAAS	*HLA-DRB1*01:02, HLA-DRB1*09:01, HLA-DRB1*10:01, HLA-DRB1*04:03, HLA-DRB1*08:02, HLA-DRB1*04:06*	21.32%	2.58%	29.64%	16.28%
Rv1800	169	SWLQRLQSIPGAASL	*HLA-DRB1*01:02, HLA-DRB1*09:01, HLA-DRB1*10:01, HLA-DRB1*04:03, HLA-DRB1*08:02, HLA-DRB1*04:06*	21.32%	2.58%	29.64%	16.28%
Rv1800	192	EAPMGVVRAVNSAIA	*HLA-DRB1*01:02, HLA-DRB1*09:01, HLA-DRB1*13:02, HLA-DRB1*08:02, HLA-DRB1*11:03, HLA-DRB1*13:03*	27.92%	7.65%	49.72%	37.62%
Rv1800	193	APMGVVRAVNSAIAA	*HLA-DRB1*01:02, HLA-DRB1*09:01, HLA-DRB1*13:02, HLA-DRB1*08:02, HLA-DRB1*11:03, HLA-DRB1*13:03*	27.92%	7.65%	49.72%	37.62%
Rv1800	194	PMGVVRAVNSAIAAN	*HLA-DRB1*01:02, HLA-DRB1*09:01, HLA-DRB1*13:02, HLA-DRB1*08:02, HLA-DRB1*11:03, HLA-DRB1*13:03, HLA-DQA1*04:01/DQB1*04:02*	27.92%	7.65%	49.72%	37.62%
Rv1800	195	MGVVRAVNSAIAANA	*HLA-DRB1*01:02, HLA-DRB1*09:01, HLA-DRB1*13:02, HLA-DRB1*08:02, HLA-DRB1*11:03, HLA-DRB1*13:03, HLA-DQA1*04:01/DQB1*04:02*	27.92%	7.65%	49.72%	37.62%
Rv1800	196	GVVRAVNSAIAANAA	*HLA-DRB1*01:02, HLA-DRB1*09:01, HLA-DQA1*01:02/DQB1*06:02, HLA-DRB1*08:02, HLA-DQA1*04:01/DQB1*04:02*	14.62%	1.79%	17.63%	11.17%

Protein	Pos	Epitope	HLA Alleles	Zim-babwe	South Africa (Black)	Ethiopia	Congo
Rv1800	257	GLYPVVVIKNLTFDS	*HLA-DRB1*01:02, HLA-DRB1*15:02, HLA-DRB1*11:04, HLA-DRB1*13:01, HLA-DRB1*11:02*	35.84%	23.96%	22.03%	32.38%
Rv1800	258	LYPVVVIKNLTFDSS	*HLA-DRB1*01:02, HLA-DRB1*15:02, HLA-DRB1*11:04, HLA-DRB1*13:01, HLA-DRB1*11:02*	35.84%	23.96%	22.03%	32.38%
Rv1800	259	YPVVVIKNLTFDSSV	*HLA-DRB1*01:02, HLA-DRB1*15:02, HLA-DRB1*11:04, HLA-DRB1*13:01, HLA-DRB1*11:02*	35.84%	23.96%	22.03%	32.38%
Rv1800	260	PVVVIKNLTFDSSVA	*HLA-DRB1*01:02, HLA-DRB1*15:02, HLA-DRB1*11:04, HLA-DRB1*13:01, HLA-DRB1*11:02*	35.84%	23.96%	22.03%	32.38%
Rv1800	261	VVVIKNLTFDSSVAQ	*HLA-DRB1*01:02, HLA-DRB1*15:02, HLA-DRB1*11:04, HLA-DRB1*13:01, HLA-DRB1*11:02*	35.84%	23.96%	22.03%	32.38%
Rv1800	302	SATISSLVMANLAAS	*HLA-DRB1*11:04, HLA-DRB1*08:04, HLA-DRB1*08:02, HLA-DRB1*13:01, HLA-DRB1*11:02*	33.41%	23.96%	16.88%	35.87%
Rv1800	303	ATISSLVMANLAASA	*HLA-DRB1*11:04, HLA-DRB1*08:04, HLA-DRB1*08:02, HLA-DRB1*13:01, HLA-DRB1*11:02*	33.41%	23.96%	16.88%	35.87%
Rv1800	304	TISSLVMANLAASAD	*HLA-DRB1*11:04, HLA-DRB1*08:04, HLA-DRB1*08:02, HLA-DRB1*13:01, HLA-DRB1*11:02*	33.41%	23.96%	16.88%	35.87%
Rv1800	305	ISSLVMANLAASADP	*HLA-DRB1*11:04, HLA-DRB1*08:04, HLA-DRB1*08:02, HLA-DRB1*13:01, HLA-DRB1*11:02*	33.41%	23.96%	16.88%	35.87%
Rv1800	306	SSLVMANLAASADPP	*HLA-DRB1*11:04, HLA-DRB1*08:04, HLA-DRB1*08:02, HLA-DRB1*13:01, HLA-DRB1*11:02*	33.41%	23.96%	16.88%	35.87%
Rv1800	307	SLVMANLAASADPPS	*HLA-DRB1*11:04, HLA-DRB1*08:04, HLA-DRB1*08:02, HLA-DRB1*13:01, HLA-DRB1*11:02*	33.41%	23.96%	16.88%	35.87%
Rv1800	308	LVMANLAASADPPSP	*HLA-DRB1*11:04, HLA-DRB1*08:04, HLA-DRB1*08:02, HLA-DRB1*13:01, HLA-DRB1*11:02*	33.41%	23.96%	16.88%	35.87%
Rv1800	381	PRYPLNFVSTLNAIA	*HLA-DRB1*11:04, HLA-DRB1*04:03, HLA-DRB1*08:02, HLA-DRB1*04:06, HLA-DRB1*04:05, HLA-DRB1*11:01, HLA-DRB1*04:01*	31.61%	0.00%	24.67%	23.79%
Rv1800	382	RYPLNFVSTLNAIAG	*HLA-DRB1*11:04, HLA-DRB1*14:04, HLA-DRB1*10:01, HLA-DRB1*04:03, HLA-DRB1*08:02, HLA-DRB1*04:06, HLA-DRB1*04:05, HLA-DRB1*11:01, HLA-DRB1*13:03, HLA-DRB1*04:01*	39.00%	0.80%	33.94%	33.37%
Rv1800	383	YPLNFVSTLNAIAGT	*HLA-DRB1*11:04, HLA-DRB1*10:01, HLA-DRB1*04:03, HLA-DRB1*08:02, HLA-DRB1*04:06, HLA-DRB1*04:05, HLA-DRB1*11:01, HLA-DRB1*13:03, HLA-DRB1*04:01*	39.00%	0.80%	33.94%	33.37%
Rv1800	384	PLNFVSTLNAIAGTY	*HLA-DRB1*11:04, HLA-DRB1*10:01, HLA-DRB1*04:03, HLA-DRB1*08:02, HLA-DRB1*04:06, HLA-DRB1*04:05, HLA-DRB1*11:01, HLA-DRB1*13:03, HLA-DRB3*02:02, HLA-DRB1*04:01*	39.00%	0.80%	33.94%	33.37%
Rv1800	385	LNFVSTLNAIAGTYY	*HLA-DRB1*11:04, HLA-DRB1*10:01, HLA-DRB1*04:03, HLA-DRB1*08:02, HLA-DRB1*04:06, HLA-DRB1*04:05, HLA-DRB1*11:01, HLA-DRB1*04:01*	37.59%	0.80%	29.42%	28.51%
Rv1800	425	TMTQYYIIRTENLPL	*HLA-DRB1*13:02, HLA-DRB1*14:01, HLA-DRB1*14:04, HLA-DRB1*04:03, HLA-DRB1*04:06, HLA-DRB1*04:05, HLA-DRB1*13:03, HLA-DRB1*07:01, HLA-DRB1*04:01*	29.10%	5.91%	77.58%	44.29%
Rv1800	426	MTQYYIIRTENLPLL	*HLA-DRB1*14:05, HLA-DRB1*08:03, HLA-DRB1*13:02, HLA-DRB1*14:01, HLA-DRB1*14:04, HLA-DRB1*10:01, HLA-DRB1*04:03, HLA-DRB1*04:06, HLA-DRB1*04:05, HLA-DRB1*11:03, HLA-DRB1*13:03, HLA-DRB1*07:01, HLA-DRB3*02:02, HLA-DRB1*04:01*	35.20%	6.68%	80.14%	48.32%
Rv1800	427	TQYYIIRTENLPLLE	*HLA-DRB1*14:05, HLA-DRB1*13:02, HLA-DRB1*14:01, HLA-DRB1*14:04, HLA-DRB1*10:01, HLA-DRB1*04:03, HLA-DRB1*04:06, HLA-DRB1*04:05, HLA-DRB1*11:03, HLA-DRB1*13:03, HLA-DRB1*07:01, HLA-DRB3*02:02, HLA-DRB1*04:01*	35.20%	6.68%	80.14%	48.32%
Rv1800	428	QYYIIRTENLPLLEP	*HLA-DRB1*14:05, HLA-DRB1*13:02, HLA-DRB1*14:01, HLA-DRB1*14:04, HLA-DRB1*04:03, HLA-DRB1*04:06, HLA-DRB1*04:05, HLA-DRB1*11:03, HLA-DRB1*13:03, HLA-DRB1*07:01, HLA-DRB3*02:02*	29.10%	5.91%	76.00%	42.63%

Protein	Pos	Epitope	HLA Alleles	Zim-babwe	South Africa (Black)	Ethiopia	Congo
Rv1800	429	YYIIRTENLPLLEPL	*HLA-DRB1*13:02, HLA-DRB1*14:01, HLA-DRB1*04:05, HLA-DRB1*11:03, HLA-DRB1*13:03, HLA-DRB3*02:02*	21.15%	5.91%	45.11%	32.33%
Rv1800	490	FPEVSPVVIADALVA	*HLA-DRB1*01:02, HLA-DRB1*15:02, HLA-DRB1*11:04, HLA-DRB1*08:04, HLA-DRB1*13:01, HLA-DRB1*11:02*	43.00%	23.96%	32.87%	41.28%
Rv1800	491	PEVSPVVIADALVAG	*HLA-DRB1*01:02, HLA-DRB1*15:02, HLA-DRB1*11:04, HLA-DRB1*08:04, HLA-DRB1*13:01, HLA-DRB1*11:02, HLA-DRB1*03:01*	53.08%	45.98%	46.91%	49.76%
Rv1800	492	EVSPVVIADALVAGT	*HLA-DRB1*01:02, HLA-DRB1*15:02, HLA-DRB1*11:04, HLA-DRB1*08:04, HLA-DRB1*13:01, HLA-DRB1*11:02, HLA-DRB1*03:01*	53.08%	45.98%	46.91%	49.76%
Rv1800	493	VSPVVIADALVAGTQ	*HLA-DRB1*01:02, HLA-DRB1*15:02, HLA-DRB1*11:04, HLA-DRB1*08:04, HLA-DRB1*13:01, HLA-DRB1*11:02, HLA-DRB1*03:01*	53.08%	45.98%	46.91%	49.76%
Rv1800	494	SPVVIADALVAGTQQ	*HLA-DRB1*01:02, HLA-DRB1*15:02, HLA-DRB1*11:04, HLA-DRB1*08:04, HLA-DRB1*13:01, HLA-DRB1*11:02, HLA-DRB1*03:01*	53.08%	45.98%	46.91%	49.76%
Rv1800	495	PVVIADALVAGTQQG	*HLA-DRB1*01:02, HLA-DRB1*15:02, HLA-DRB1*11:04, HLA-DRB1*08:04, HLA-DRB1*13:01, HLA-DRB1*11:02*	43.00%	23.96%	32.87%	41.28%
Rv1800	496	VVIADALVAGTQQGI	*HLA-DRB1*01:02, HLA-DRB1*15:02, HLA-DRB1*11:04, HLA-DRB1*08:04, HLA-DRB1*13:01, HLA-DRB1*11:02*	43.00%	23.96%	32.87%	41.28%
Rv1800	584	NAALTIVPSYNIHLF	*HLA-DRB1*01:02, HLA-DRB1*08:02, HLA-DRB1*11:03, HLA-DRB1*15:03, HLA-DRB1*15:01*	38.53%	0.00%	36.86%	6.78%
Rv1800	630	AAGGLQLLIIISAGR	*HLA-DRB1*11:04, HLA-DRB1*08:04, HLA-DRB1*11:01, HLA-DRB1*15:01, HLA-DRB5*01:01*	35.84%	0.00%	20.42%	29.61%
Rv1800	631	AGGLQLLIIISAGRT	*HLA-DRB1*11:04, HLA-DRB1*08:04, HLA-DRB1*11:01, HLA-DRB1*15:01, HLA-DRB5*01:01*	35.84%	0.00%	20.42%	29.61%
Rv1800	632	GGLQLLIIISAGRTI	*HLA-DRB1*01:02, HLA-DRB1*11:04, HLA-DRB1*08:04, HLA-DRB1*12:01, HLA-DRB1*11:01, HLA-DRB1*15:01, HLA-DRB5*01:01, HLA-DRB1*12:02*	53.62%	0.00%	36.50%	42.16%
Rv1800	633	GLQLLIIISAGRTIA	*HLA-DRB1*01:02, HLA-DRB1*11:04, HLA-DRB1*08:04, HLA-DRB1*12:01, HLA-DRB1*11:01, HLA-DRB1*11:03, HLA-DRB1*15:03, HLA-DRB1*15:01, HLA-DRB5*01:01, HLA-DRB1*12:02*	72.12%	0.00%	50.18%	42.16%
Rv1800	634	LQLLIIISAGRTIAN	*HLA-DRB1*01:02, HLA-DRB1*11:04, HLA-DRB1*08:04, HLA-DRB1*12:01, HLA-DRB1*11:01, HLA-DRB1*11:03, HLA-DRB1*15:03, HLA-DRB1*15:01, HLA-DRB5*01:01, HLA-DRB1*12:02*	72.12%	0.00%	50.18%	42.16%
Rv1800	635	QLLIIISAGRTIAND	*HLA-DRB1*01:02, HLA-DRB1*11:03, HLA-DRB1*15:01, HLA-DRB5*01:01, HLA-DRB1*12:02*	13.51%	0.00%	21.55%	6.78%
Rv2356c	1	VVNFSVLPPEINSGR	*HLA-DRB1*13:01, HLA-DRB1*15:02, HLA-DRB1*11:01, HLA-DRB1*11:02, HLA-DRB1*11:04*	48.45%	23.96%	9.16%	40.12%
Rv2356c	110	DPVVVAANRSAFVQL	*HLA-DRB1*11:02, HLA-DRB1*08:02, HLA-DRB1*11:03, HLA-DRB1*09:01, HLA-DRB1*03:01*	24.48%	27.07%	17.83%	25.44%
Rv2356c	111	PVVVAANRSAFVQLV	*HLA-DRB1*11:02, HLA-DRB1*08:02, HLA-DRB1*11:03, HLA-DRB1*09:01, HLA-DRB1*03:01*	24.48%	27.07%	17.83%	25.44%
Rv2356c	118	RSAFVQLVLSNVFGQ	*HLA-DRB1*13:01, HLA-DRB1*08:04, HLA-DRB1*11:01, HLA-DRB1*04:05, HLA-DRB1*11:02, HLA-DRB1*11:04*	57.75%	23.96%	29.94%	50.04%
Rv2356c	119	SAFVQLVLSNVFGQN	*HLA-DRB1*13:01, HLA-DRB1*08:04, HLA-DRB1*11:01, HLA-DRB1*04:05, HLA-DRB1*11:02, HLA-DRB1*11:04*	57.75%	23.96%	29.94%	50.04%

Protein	Pos	Epitope	HLA Alleles	Zim-babwe	South Africa (Black)	Ethiopia	Congo
Rv2356c	123	QLVLSNVFGQNAPAI	*HLA-DRB1*13:01, HLA-DRB1*08:04, HLA-DRB1*01:02, HLA-DRB1*11:02, HLA-DRB1*11:04*	43.00%	23.96%	32.87%	41.28%
Rv2356c	124	LVLSNVFGQNAPAIA	*HLA-DRB1*13:01, HLA-DRB1*08:04, HLA-DRB1*01:02, HLA-DRB1*11:02, HLA-DRB1*11:04*	43.00%	23.96%	32.87%	41.28%
Rv2356c	128	NVFGQNAPAIAAAEA	*HLA-DQA1*04:01/DQB1*04:02, HLA-DRB1*01:02, HLA-DQA1*01:02/DQB1*06:02, HLA-DRB1*09:01, HLA-DQA1*03:01/DQB1*03:02*	14.62%	1.79%	17.63%	11.17%
Rv2608	343	LAAGNEVVVFGTSQS	*HLA-DRB1*11:04, HLA-DRB1*13:01, HLA-DRB1*11:02, HLA-DRB1*15:02, HLA-DRB1*08:04*	33.41%	23.96%	16.88%	35.87%
Rv2608	344	AAGNEVVVFGTSQSA	*HLA-DRB1*11:04, HLA-DRB1*13:01, HLA-DRB1*11:02, HLA-DRB1*15:02, HLA-DRB1*08:04*	33.41%	23.96%	16.88%	35.87%
Rv2608	345	AGNEVVVFGTSQSAT	*HLA-DRB1*11:04, HLA-DRB1*13:01, HLA-DRB1*11:02, HLA-DRB1*15:02, HLA-DRB1*08:04*	33.41%	23.96%	16.88%	35.87%
Rv2608	346	GNEVVVFGTSQSATI	*HLA-DRB1*11:04, HLA-DRB1*13:01, HLA-DRB1*11:02, HLA-DRB1*15:02, HLA-DRB1*08:04, HLA-DRB1*07:01*	40.71%	23.96%	50.76%	45.89%
Rv2608	347	NEVVVFGTSQSATIA	*HLA-DRB1*11:04, HLA-DRB1*08:02, HLA-DRB1*13:01, HLA-DRB1*11:02, HLA-DRB1*15:02, HLA-DRB1*04:01, HLA-DRB1*08:04*	33.41%	23.96%	19.84%	37.62%
Rv2608	348	EVVVFGTSQSATIAT	*HLA-DRB1*11:04, HLA-DRB1*13:01, HLA-DRB1*11:02, HLA-DRB1*15:02, HLA-DRB1*04:01, HLA-DRB1*08:04*	33.41%	23.96%	19.84%	37.62%
Rv2608	349	VVVFGTSQSATIATF	*HLA-DRB1*11:04, HLA-DRB1*13:01, HLA-DRB1*11:02, HLA-DRB1*09:01, HLA-DRB1*15:02, HLA-DRB1*04:01, HLA-DRB1*08:04*	35.84%	25.52%	19.84%	41.20%
Rv2608	359	TIATFEMRYLQSLPA	*HLA-DRB1*11:04, HLA-DRB1*01:02, HLA-DRB1*11:01, HLA-DRB1*13:01, HLA-DRB1*11:02, HLA-DRB1*04:06, HLA-DRB1*15:02, HLA-DRB1*04:03, HLA-DRB1*08:04*	62.67%	23.96%	42.85%	53.31%
Rv2608	360	IATFEMRYLQSLPAH	*HLA-DRB1*11:04, HLA-DRB1*01:02, HLA-DRB1*11:01, HLA-DRB1*13:01, HLA-DRB1*11:02, HLA-DRB1*04:06, HLA-DRB1*10:01, HLA-DRB1*15:02, HLA-DRB1*04:03, HLA-DRB1*08:04*	67.05%	24.66%	46.97%	56.99%
Rv2608	361	ATFEMRYLQSLPAHL	*HLA-DRB1*11:04, HLA-DRB1*01:02, HLA-DRB1*14:05, HLA-DRB1*11:01, HLA-DRB1*01:01, HLA-DRB1*13:01, HLA-DRB1*11:02, HLA-DRB1*04:06, HLA-DRB1*16:02, HLA-DRB1*13:03, HLA-DRB1*15:01, HLA-DRB1*09:01, HLA-DRB3*02:02, HLA-DRB1*15:04, HLA-DRB1*14:01, HLA-DRB1*10:01, HLA-DRB1*15:02, HLA-DRB1*15:03, HLA-DRB1*04:03, HLA-DRB1*04:01, HLA-DRB1*04:05, HLA-DRB1*08:04, HLA-DRB1*14:04, HLA-DRB1*07:01*	91.99%	26.21%	90.95%	77.47%
Rv2608	362	TFEMRYLQSLPAHLR	*HLA-DRB1*11:04, HLA-DRB1*01:02, HLA-DRB1*14:05, HLA-DRB1*11:01, HLA-DRB1*01:01, HLA-DRB1*13:01, HLA-DRB1*11:02, HLA-DRB1*04:06, HLA-DRB1*16:02, HLA-DRB1*13:03, HLA-DRB1*15:01, HLA-DRB1*09:01, HLA-DRB3*02:02, HLA-DRB1*15:04, HLA-DRB1*14:01, HLA-DRB1*10:01, HLA-DRB1*03:02, HLA-DRB1*15:02, HLA-DRB1*15:03, HLA-DRB1*04:03, HLA-DRB1*04:01, HLA-DRB5*01:01, HLA-DRB1*04:05, HLA-DRB1*16:01, HLA-DRB1*08:04, HLA-DRB1*14:04, HLA-DRB1*07:01*	95.29%	40.86%	90.95%	80.62%

Protein	Pos	Epitope	HLA Alleles	Zim-babwe	South Africa (Black)	Ethiopia	Congo
Rv2608	363	FEMRYLQSLPAHLRP	*HLA-DRB1*11:04, HLA-DRB1*01:02, HLA-DRB1*14:05, HLA-DRB1*11:01, HLA-DRB1*01:01, HLA-DRB1*13:01, HLA-DRB1*11:02, HLA-DRB1*04:06, HLA-DRB1*16:02, HLA-DRB1*13:03, HLA-DRB1*15:01, HLA-DRB1*09:01, HLA-DRB3*02:02, HLA-DRB1*15:04, HLA-DRB1*14:01, HLA-DRB1*10:01, HLA-DRB1*03:02, HLA-DRB1*15:02, HLA-DRB1*15:03, HLA-DRB1*04:03, HLA-DRB1*04:01, HLA-DRB5*01:01, HLA-DRB1*04:05, HLA-DRB1*16:01, HLA-DRB1*08:04, HLA-DRB1*14:04, HLA-DRB1*07:01*	95.29%	40.86%	90.95%	80.62%
Rv2608	364	EMRYLQSLPAHLRPG	*HLA-DRB1*11:04, HLA-DRB1*01:02, HLA-DRB1*14:05, HLA-DRB1*11:01, HLA-DRB1*01:01, HLA-DRB1*13:01, HLA-DRB1*11:02, HLA-DRB1*04:06, HLA-DRB1*16:02, HLA-DRB1*13:03, HLA-DRB1*15:01, HLA-DRB1*09:01, HLA-DRB3*02:02, HLA-DRB1*15:04, HLA-DRB1*14:01, HLA-DRB1*10:01, HLA-DRB1*03:02, HLA-DRB1*15:02, HLA-DRB1*15:03, HLA-DRB1*04:03, HLA-DRB5*01:01, HLA-DRB1*16:01, HLA-DRB1*08:04, HLA-DRB1*07:01*	94.29%	40.86%	86.33%	78.64%
Rv2608	365	MRYLQSLPAHLRPGL	*HLA-DRB1*11:04, HLA-DRB1*01:02, HLA-DRB1*14:05, HLA-DRB1*11:01, HLA-DRB1*01:01, HLA-DRB1*13:01, HLA-DRB1*11:02, HLA-DRB1*04:06, HLA-DRB1*16:02, HLA-DRB1*09:01, HLA-DRB3*02:02, HLA-DRB1*10:01, HLA-DRB1*03:02, HLA-DRB1*15:02, HLA-DRB1*04:03, HLA-DRB5*01:01, HLA-DRB1*16:01, HLA-DRB1*08:04*	78.47%	40.86%	46.97%	67.48%
Rv2608	366	RYLQSLPAHLRPGLD	*HLA-DRB1*09:01, HLA-DRB3*02:02, HLA-DRB1*10:01, HLA-DRB1*03:02, HLA-DRB5*01:01*	22.21%	19.54%	5.48%	15.18%
Rv2608	438	KYPLNVFATANAIAG	*HLA-DRB1*08:02, HLA-DRB1*11:02, HLA-DRB1*04:06, HLA-DRB1*09:01, HLA-DRB1*04:03*	11.83%	1.79%	8.57%	15.18%
Rv2608	478	PDVLTTYILLPSQDL	*HLA-DRB1*01:02, HLA-DRB1*01:01, HLA-DPA1*02:01/DPB1*01:01, HLA-DPA1*03:01/DPB1*04:02, HLA-DPA1*01:03/DPB1*02:01, HLA-DRB5*01:01, HLA-DRB1*04:05, HLA-DRB1*08:04*	27.58%	0.00%	37.35%	23.17%
Rv2608	479	DVLTTYILLPSQDLP	*HLA-DRB1*01:02, HLA-DRB1*01:01, HLA-DPA1*02:01/DPB1*01:01, HLA-DPA1*03:01/DPB1*04:02, HLA-DRB1*10:01, HLA-DRB5*01:01, HLA-DRB1*04:05, HLA-DRB1*08:04*	33.74%	0.80%	41.68%	27.92%
Rv2608	480	VLTTYILLPSQDLPL	*HLA-DRB1*01:02, HLA-DRB1*01:01, HLA-DRB1*04:06, HLA-DRB1*10:01, HLA-DRB1*04:03, HLA-DRB5*01:01, HLA-DRB1*04:05, HLA-DRB1*08:04*	33.74%	0.80%	47.77%	27.92%
Rv2608	481	LTTYILLPSQDLPLL	*HLA-DRB1*01:02, HLA-DRB1*01:01, HLA-DRB1*04:06, HLA-DRB1*10:01, HLA-DRB1*04:03, HLA-DRB5*01:01, HLA-DRB1*04:05, HLA-DRB1*08:04*	33.74%	0.80%	47.77%	27.92%
Rv2608	482	TTYILLPSQDLPLLV	*HLA-DRB1*01:02, HLA-DRB1*01:01, HLA-DRB1*04:06, HLA-DRB1*10:01, HLA-DRB1*04:03, HLA-DRB5*01:01, HLA-DRB1*04:05*	26.04%	0.80%	38.16%	18.10%
Rv2608	489	SQDLPLLVPLRAIPL	*HLA-DRB1*11:04, HLA-DRB1*01:02, HLA-DRB1*11:01, HLA-DRB1*15:01, HLA-DRB1*15:02*	38.22%	0.00%	25.45%	25.95%
Rv2608	490	QDLPLLVPLRAIPLL	*HLA-DRB1*11:04, HLA-DRB1*01:02, HLA-DRB1*11:01, HLA-DRB1*15:01, HLA-DRB1*15:02*	38.22%	0.00%	25.45%	25.95%
Rv2608	491	DLPLLVPLRAIPLLG	*HLA-DRB1*11:04, HLA-DRB1*01:02, HLA-DRB1*11:01, HLA-DRB1*15:01, HLA-DRB1*15:02*	38.22%	0.00%	25.45%	25.95%
Rv2608	492	LPLLVPLRAIPLLGN	*HLA-DRB1*11:04, HLA-DRB1*01:02, HLA-DRB1*11:01, HLA-DRB1*15:01, HLA-DRB1*15:02*	38.22%	0.00%	25.45%	25.95%

Protein	Pos	Epitope	HLA Alleles	Zim-babwe	South Africa (Black)	Ethiopia	Congo
Rv3533c	10	LNSLRMFTGAGSAPM	*HLA-DRB1*08:04, HLA-DRB1*15:02, HLA-DRB1*09:01, HLA-DRB1*04:01, HLA-DRB1*13:01, HLA-DRB1*11:02, HLA-DRB1*01:02, HLA-DRB1*11:04*	45.24%	25.52%	35.54%	46.37%
Rv3533c	105	VHPLLVAANRNAFAQ	*HLA-DRB1*08:02, HLA-DRB1*11:03, HLA-DRB1*03:01, HLA-DRB1*13:01, HLA-DRB1*13:03*	30.11%	45.98%	26.01%	28.07%
Rv3533c	106	HPLLVAANRNAFAQL	*HLA-DRB1*08:02, HLA-DRB1*11:03, HLA-DRB1*03:01, HLA-DRB1*13:01, HLA-DRB1*03:02, HLA-DRB1*13:03, HLA-DRB3*02:02*	40.71%	58.40%	26.01%	32.83%
Rv3533c	107	PLLVAANRNAFAQLV	*HLA-DRB1*08:02, HLA-DRB1*11:03, HLA-DRB1*03:01, HLA-DRB1*13:01, HLA-DRB1*13:03, HLA-DRB3*02:02*	30.11%	45.98%	26.01%	28.07%
Rv3533c	11	NSLRMFTGAGSAPML	*HLA-DRB1*08:04, HLA-DRB1*15:02, HLA-DRB1*09:01, HLA-DRB1*04:01, HLA-DRB1*13:01, HLA-DRB1*11:02, HLA-DRB1*01:02, HLA-DRB1*11:04*	45.24%	25.52%	35.54%	46.37%
Rv3533c	12	SLRMFTGAGSAPMLA	*HLA-DRB1*08:04, HLA-DRB1*15:02, HLA-DRB1*10:01, HLA-DRB1*04:01, HLA-DRB1*13:01, HLA-DRB1*11:02, HLA-DRB1*01:02, HLA-DRB1*11:04*	48.45%	24.66%	39.92%	47.03%
Rv3533c	13	LRMFTGAGSAPMLAA	*HLA-DRB1*08:04, HLA-DRB1*15:02, HLA-DRB1*10:01, HLA-DRB1*04:01, HLA-DRB1*13:01, HLA-DRB1*11:02, HLA-DRB1*01:02, HLA-DRB1*11:04*	48.45%	24.66%	39.92%	47.03%
Rv3533c	151	MVGYHSGASAAAEQL	*HLA-DRB1*08:04, HLA-DRB1*15:02, HLA-DQA1*04:01/DQB1*04:02, HLA-DQA1*03:01/DQB1*03:02, HLA-DQA1*05:01/DQB1*03:01*	8.99%	0.00%	12.33%	10.89%
Rv3533c	168	FQQALQQLPNLGIGN	*HLA-DRB1*15:02, HLA-DRB1*04:03, HLA-DRB1*10:01, HLA-DRB1*04:06, HLA-DRB1*01:02*	18.64%	0.80%	29.64%	12.02%
Rv3533c	169	QQALQQLPNLGIGNI	*HLA-DRB1*15:02, HLA-DRB1*04:03, HLA-DRB1*10:01, HLA-DRB1*04:06, HLA-DRB1*01:02*	18.64%	0.80%	29.64%	12.02%
Rv3533c	7	PPELNSLRMFTGAGS	*HLA-DRB1*08:04, HLA-DRB1*15:02, HLA-DRB1*11:01, HLA-DRB1*09:01, HLA-DRB1*13:01, HLA-DRB1*11:02, HLA-DRB1*01:02, HLA-DRB1*11:04*	64.48%	25.52%	36.48%	56.40%
Rv3533c	74	GWLSAAAARAAGAAA	*HLA-DQA1*04:01/DQB1*04:02, HLA-DRB1*09:01, HLA-DQA1*05:01/DQB1*03:01, HLA-DQA1*01:02/DQB1*06:02, HLA-DRB1*01:01*	6.88%	1.79%	0.00%	8.80%
Rv3533c	8	PELNSLRMFTGAGSA	*HLA-DRB1*08:04, HLA-DRB1*15:02, HLA-DRB1*11:01, HLA-DRB1*09:01, HLA-DRB1*13:01, HLA-DRB1*11:02, HLA-DRB1*01:02, HLA-DRB1*11:04*	64.48%	25.52%	36.48%	56.40%
Rv3533c	9	ELNSLRMFTGAGSAP	*HLA-DRB1*08:04, HLA-DRB1*15:02, HLA-DRB1*09:01, HLA-DRB1*13:01, HLA-DRB1*11:02, HLA-DRB1*01:02, HLA-DRB1*11:04*	45.24%	25.52%	32.87%	44.75%